Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

International Journal of Information Technology & Decision Making
Vol. 18, No. 3 (2019) 747–791
© World Scientific Publishing Company
DOI: 10.1142/S0219622019500172

World Scientific
www.worldscientific.com

Conversion Uplift in E-Commerce: A Systematic Benchmark of Modeling Strategies

Robin Gubela, Artem Bequé and Stefan Lessmann

School of Business and Economics
Humboldt-University of Berlin
Unter den Linden 6, 10099
Berlin, Germany

Fabian Gebert

Data Science Department
Akanoo GmbH, Mittelweg 121
20148 Hamburg, Germany

Published 3 May 2019

Uplift modeling combines machine learning and experimental strategies to estimate the differential effect of a treatment on individuals' behavior. The paper considers uplift models in the scope of marketing campaign targeting. Literature on uplift modeling strategies is fragmented across academic disciplines and lacks an overarching empirical comparison. Using data from online retailers, we fill this gap and contribute to literature through consolidating prior work on uplift modeling and systematically comparing the predictive performance and utility of available uplift modeling strategies. Our empirical study includes three experiments in which we examine the interaction between an uplift modeling strategy and the underlying machine learning algorithm to implement the strategy, quantify model performance in terms of business value and demonstrate the advantages of uplift models over response models, which are widely used in marketing. The results facilitate making specific recommendations how to deploy uplift models in e-commerce applications.

Keywords: E-commerce analytics; machine learning; uplift modeling; real-time targeting.

1. Introduction

The meteoric rise of electronic commerce and continuous growth in internet adoption leads many business organizations to adopt digital channels for expanding their market presence.[1] For example, the Digital Commerce 360[a] report predicts that electronic commerce will be 17% of the U.S. retail sales by 2022. However, online marketplaces also increase competition and pose several challenges. For example,

[a]Data from the official website retrieved on 22.09.2017 (https://www.digitalcommerce360.com/2017/08/09/e-commerce-grow-17-us-retail-sales-2022/).

lower search cost increase price competition and diminish seller profits.[2] Similarly, negative consumer reviews adversely affect company reputation and may cause financial losses.[3] To cope with these challenges, companies execute digital marketing strategies to achieve their business objectives and to survive in a challenging business environment.

Digital marketing uses analytic methods to extract relevant insights from massive amounts of data and to drive the company toward growth and profitability. Empirical marketing decision models support all stages of a customer lifecycle including acquisition, development, and retention management.[4,5] Digital marketing is most successful when it is personalized and well-targeted.[6] To target marketing communication, marketers use response models that predict customer behavior[7,8] and, in particular, the likelihood of customers to respond to a marketing offer.[9] There are many examples of response modeling in digital marketing. Use cases include targeting customers with email-based digital coupons,[10] a dynamic adaptation of websites to infer user intentions,[11] or prediction of the success of social media initiatives.[12]

In the context of marketing campaign planning, response models suffer a limitation. They fail to distinguish different customer segments.[13,14] To illustrate this, consider a marketing campaign aimed at soliciting digital coupons.[15] To efficiently allocate the marketing budget available for the campaign, a marketer wants to offer a coupon only to those customers who do not buy without such price reduction.[16] Response models ignore the causal link between the marketing action and customer response.[17] Instead, they recommend targeting customers with highest likelihood to buy. Such targeting inevitably leads to soliciting customers who would also buy without an incentive and thus wastes marketing resources.[18] Uplift models add the element of causality that response models miss. They identify customers who buy because of a marketing action and enable better campaign targeting.[19,20] More specifically, uplift models identify customers who are likely to change their behavior in response to a marketing message.[21] This is equivalent to modeling the differential (i.e., causal) effect of a marketing incentive on customer behavior.

Several approaches for uplift modeling have appeared in the literature (see Ref. 22 for a recent survey). This paper focuses on uplift modeling strategies that work together with existing algorithms from supervised machine learning. The important advantage of corresponding strategies for corporate practice is that they facilitate predicting uplift and overcome the limitations of response models, while avoiding the need to invest in new technology. Leveraging supervised machine learning, a technology widely available and used in corporate environments,[23] corresponding uplift modeling strategies are relatively easy to adopt.

As we detail in the review of related literature, previous work on uplift models does not emphasize the advantage of uplift modeling strategies to avoid large upfront investments (for example into new software) through reusing supervised machine learning. Moreover, little attempt has been made to systematically explore their potential for conversion modeling. The need for a comprehensive benchmark emerges

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

due to the available uplift modeling strategies coming from different academic disciplines. Furthermore, the few papers that employ uplift modeling strategies consider only a small set of learning algorithms — typically only one — and do not examine interactions between different classification algorithms and uplift modeling strategies. Consequently, guidance showing which classifiers work well with which uplift modeling strategy is missing. The goal of the paper is to fill these research gaps. To achieve this, we integrate previous literature on uplift modeling, evaluate the effectiveness of alternative uplift modeling strategies for campaign planning, and examine the degree to which this effectiveness depends on the learning algorithm to implement the modeling strategy.

In pursuing its objective, the paper makes the following contributions. First, we consolidate the state-of-the-art in uplift modeling. The comprehensive literature examination helps us understand and clarify the conceptual differences between different approaches and, through studying different streams of research, provides an update on modern uplift modeling research. Second, we empirically evaluate the performance of uplift modeling strategies for conversion modeling through large-scale experimentation. In particular, we benchmark several strategies across numerous datasets of different product lines from multiple geographies. The benchmark experiment provides a reference point for other academics and practitioners in campaign planning and uplift modeling. In total, our empirical study includes three experiments. For the first experiment, we consider multiple machine learning algorithms for the experimentation that we pair with each uplift modeling strategy in a full-factorial setup. Thus, we shed light on the interactions between uplift modeling strategies and underlying learning algorithms, and provide specific recommendations on their relative suitability. Based on the benchmark results, in the second experiment, we quantify the degree to which targeting marketing campaigns using uplift modeling increases business value. That is, we explain which strategy contributes most (least) to business value. To clarify differences in performance between response and uplift modeling, we compare the former with the latter in a third and final experiment.

2. Conversion Modeling using Uplift Versus Response Models

The term conversion modeling encompasses a set of marketing decision support models that estimate the probability of customers to react toward a marketing action in a way intended by the marketer. The goal of developing a conversion model is to allocate marketing resources efficiently. For example, marketers use conversion models to identify the most suitable channel to contact a customer in multi-channel advertising,[24] to select responsive customers for email surveys,[25] or, more generally, to inform targeting decisions.[18,26] These examples illustrate how conversion modeling finds broad application in marketing and e-commerce to anticipate customer behavior and to increase conversion rates.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

We distinguish conversion models into response and uplift models. Response models rely on supervised classification algorithms (hereafter, base learners), which estimate a functional relationship between a binary class label (i.e., response versus no response) and a set of explanatory variables that characterize customers. Such variables often include demographic, behavioral, and attitudinal information or, more generally, any piece of information an analyst believes to be possibly linked to customers' response behavior.

To target campaigns using a response model, candidate recipients are ordered according to model-estimated conversion probabilities and a fraction of the top-ranked recipients is contacted, whereby the size of the target group depends on the available budget and/or other business considerations. Uplift models do not predict conversion probabilities. Their objective is to predict how much the campaign changes the conversion probabilities of individual customers.[17,27] The important implication for campaign planning is that an uplift model will recommend a target group of *persuadable* customers, whose conversion probability raises because of the campaign. The target group recommended by a response model, on the other hand, will consist of *responsive* customers who may be influenced by the campaign or not. More formally, an uplift model estimates a conditional average treatment effect[28] and establishes a causal link between the marketing action and how it alters customer behavior. Causality is crucial in campaign targeting to maximize the efficiency of resource utilization. Marketing budget should be spent on those customers where it increases conversion probabilities the most. Table 1 further elaborates on the connection between the action and the behavioral change it induces by distinguishing four groups of customers. Without loss of generality, we assume in the following that a marketing campaign aims at direct selling. Hence, successful conversion implies that a customer purchases the offered item. We denote corresponding customers as buyers. Instead of campaign, we use the term treatment, which is more general than campaign and used in the econometrics literature on causal inference.[29] A treated customer receives the marketing action (e.g., email-based digital coupon).

According to Table 1, customers classified as *sure things* buy regardless of the treatment while *lost causes* never buy. Clearly, contacting these groups with a marketing message is a waste of resources. Even worse, the effect of a treatment is detrimental for customers in the *do-not-disturbs* group. Their conversion probability decreases when being treated. Last, the *persuadables* buy if being treated and refrain from buying otherwise. This means they buy because of the treatment and are thus the only group worth considering in targeted marketing actions.[17] Targeting

Table 1. Customer types as per uplift modeling.

		Buyer without treatment	
		Yes	No
Buyer with treatment	Yes	Sure Things	Persuadables
	No	Do-Not-Disturbs	Lost Causes

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

persuadables allows marketing managers to maximize the incremental number of purchases which implies an efficient use of marketing resources.

Table 1 also reveals a conceptual difference between response and uplift models. Response models require a labeled dataset of customers the actual buying behavior of whom is known from a past campaign. This is the standard setting in supervised learning. In addition to a target label, developing an uplift model also requires data from two groups of customers, a treatment group who received the marketing action and a control group who did not. Random trials, pilot campaigns, or A/B tests are common instruments to obtain corresponding data. Subsequently, a straightforward way to develop an uplift model, called the two-model uplift method, involves estimating two classification models from the treatment and control group data, respectively. To estimate conversion uplift as the difference in the predicted purchase probabilities with and without treatment, predictions for new customers (e.g., potential recipients of an upcoming campaign) are subtracted.[30] A formal derivation of this approach follows in Sec. 4 of this paper.

3. Prior Work in Uplift Modeling

In general, uplift models estimate the probability of occurrence of an event of interest through relating the event to a set of explanatory variables. A crucial difference to ordinary regression models or supervised machine learning is that uplift models aim at estimating how the probability of the event changes with specific actions. Estimating how treatment with a certain medication changes the survival probability of a patient exemplifies this approach in a medical application context. Here, the action is to apply the medication or refrain from doing so.[31] Another example arises in marketing where a marketer is interested to estimate how actions in the form of marketing messages (newsletters, telephone calls, etc.) alter the purchase behavior of customers.[32] These examples hint at the variety of applications for uplift models. The methodology underlying such models, however, is the same, which explains why prior work on uplift models spreads across different academic disciplines.

In terms of methodology, previous work on uplift models splits into three streams. The first stream comprises studies that develop uplift models using machine learning algorithms.[31,33,34] We use the term *uplift modeling strategy* to refer to corresponding approaches because they embed a conventional learning algorithm into an overall modeling framework that facilitates predicting uplift. The second stream of literature develops new learning algorithms to predict uplift.[35–37] We summarize corresponding approaches using the umbrella term *uplift algorithm*. Finally, the development of an uplift model is only one step in an overall modeling process. The third stream of research includes studies that focus on process steps preceding uplift model development such as feature selection or variable importance assessment, as employed by Ref. 38 for instance. Similarly, some studies concentrate on tasks that follow uplift model building. Nassif *et al.*[39] exemplify this approach through proposing new

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

evaluation measures for uplift model assessment. Table 2 summarizes previous work on uplift models along four dimensions: central focus of the study, uplift literature stream, experiments, industry/science, and data origin.

Table 2 reveals that much previous research is directed toward developing uplift algorithms. Corresponding works often draw inspiration from decision trees and modify algorithms for tree induction so as to predict uplift. Early work of Ref. 30 introduces tree-based uplift algorithms. Aiming at classifying recipients of a direct marketing campaign into buyers and non-buyers, their idea was to alter the splitting criterion, which governs tree growing, in such a way that it maximizes the difference between the response rate of customers in the treatment and control group. An explicit consideration of the treatment and control group alongside the class variable is the main difference to ordinary classification trees, which consider only the class variable when inducing splits. Several later studies employ a similar approach and propose improved ways to induce uplift trees. Examples include Ref. 40, who extend the X^2 criterion of the CHAID algorithm to accommodate uplift, Chickering and Heckerman,[41] who propose an approach to grow uplift trees so as to maximize expected profits, or Rzepakowski and Jaroszewicz[36] who further elaborate on tree induction through maximization of treatment and control group class distributions, and introduce novel splitting criteria based on conditional divergence. The tree-based uplift algorithm of Radcliffe and Surry[42] assesses the statistical significance of the differences among class probabilities between treatment and control group observations. Guelman[43] and Guelman, Guillén & Pérez-Marín[35] propose uplift random forest, which they derive from embedding conditional inference trees and other uplift trees in an ensemble framework. Specifically, they mimic the original random forest classifier and combine bagging with random subspace to ensemble member (uplift) models.[44] Sołtys *et al.*[45] systematize existing and contribute new uplift ensemble methods and evaluate them in marketing and medical applications.

Table 2 also illustrates that relatively few studies concentrate on uplift modeling strategies. Lo[33] as well as Tian *et al.*[34] introduce modeling strategies based on transformed data input spaces to facilitate uplift predictions. Pursuing the same goal, Jaskowski and Jaroszewicz[31] propose a methodology to modify the data output space (i.e., response variable). Shaar *et al.*[46] refer to disturbance effects of uplift models that limit prediction reliability. To cope with these effects, authors combine diverse uplift modeling strategies, including the uplift model of Lai *et al.*[47] and reflective uplift modeling in a weighted procedure to derive a pessimistic uplift score. Building on the ideas by Lai *et al.*,[47] Kane *et al.*[21] introduce a generalized weighting procedure of class probabilities.

In a benchmarking experiment, Kane *et al.*[21] empirically compare some of the above strategies. Devriendt *et al.*[22] also contrast alternative uplift models amongst which they consider uplift modeling strategies. However, as both studies exemplify, prior literature on uplift modeling strategies considers a relatively small set of supervised learning algorithms. Irrespective of the development of an uplift modeling strategy or uplift algorithm, studies generally employ a single base learner without

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 2. Prior work in uplift modeling.

Study	Main topic	Research stream	Experiment	Industry/Science	Data origin
Cai *et al.*[52]	Two-stage estimation procedure for treatment differences for HIV-infected patients	Uplift algorithm	Two treatments: (1) Therapy based on drug combination (zidovudine, lamivudine); (2) Therapy based on drug combination (zidovudine, lamivudine, indinavir)	Clinical trials	Licensed open-source real-world data (AIDS Clinical Trials Group; see study ACTG 320[57]
Chickering and Heckerman[41]	Greedy decision-tree learning algorithms (FORCE versus NORMAL)	Uplift algorithm	Mail advertisement for MSN subscription	Software	Private real-world data (anonymized authority)
Devriendt *et al.*[22]	Literature survey and empirical analysis of uplift models for marketing decision support	Specific task	Four treatments: (1) Marketing of insurances; (2) Email marketing; (3) Catalog mailing; (4) Retention marketing	Various	R-package information*; open-source real-world data[58]; data from an Udemy online course**; private real-world data from a retention program
Dost *et al.*[32]	Willingness-to-pay (WTP) range-based targeting approach	Uplift algorithm	Four experiments: (1) Discount offer; (2) WOM (T1), visual (T2), information (T3); (3) Discount (T1), guarantee (T2); (4) Participation	Various	Surveys in different settings; participants from Amazon Mechanical Turk; students from a German university; consumers from an agency panel
Guelman[43]	Personalized treatment learning problem, uplift random forest and uplift causal conditional inference forest	Uplift algorithm	Email promotion to buy a certain product at a bank	Financial services	Private real-world data (anonymized authority)

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 2. (*Continued*)

Study	Main topic	Research stream	Experiment	Industry/Science	Data origin
Guelman *et al.*[53]	Uplift random forests	Uplift algorithm	Two treatments: (1) Letter (retention); (2) Letter plus outbound courtesy call (retention)	Insurance	Private real-world data (anonymized authority)
Guelman *et al.*[54]	Causal conditional inference trees in personalized treatment learning	Uplift algorithm	Direct mail campaign (cross-selling)	Insurance	Private real-world data (anonymized authority)
Guelman *et al.*[35]	Uplift random forests	Uplift algorithm	Information letter plus courtesy call (as one treatment)	Insurance	Private real-world data (anonymized authority)
Hansen and Bowers[55]	Stratification to balance the distributions of pretreatment variables	Specific task	Especially: GOTV field experiment (GOTV messages: personal visit, phone call, mailing) and simulation studies	Social and political sciences	Get-Out-The-Vote (GOTV) field experiment[59]
Hansotia and Rukstales[19]	Concept of uplift tree-based approaches	Uplift algorithm	—	—	—
Hansotia and Rukstales[19] (2002b)	CHAID decision tree with $\Delta\Delta$P split criterion	Uplift algorithm	Mail promotion ($10 off a purchase of at least $100 basket value)	Holiday retail	Private real-world data (anonymized authority)
Hua (2016)	Uplift random forests in capital market research with focus on results of embedded variable selection procedure	Specific task	—	Financial services	Licensed open-source real-world data (different data sources)

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 2. (*Continued*)

Study	Main topic	Research stream	Experiment	Industry/Science	Data origin
Imai and Ratkovic (2013)	Estimation of heterogeneous treatment effects as a variable selection problem with modified support vector machines	Specific task	GOTV field experiment (GOTV messages: personal visit, phone call, mailing) and simulation studies	Social and political sciences	GOTV field experiment[59]
Jaroszewicz and Rzepakowski[20]	Uplift modeling for survival analysis	Uplift algorithm	Chemotherapy against colon cancer Treatment 1: Therapy with Levamisole Treatment 2: Therapy with Levamisole plus 5-Fluorouracil	Clinical trials	Open-source real-world data[60]
Jaroszewicz and Zaniewicz[50]	Uplift support vector machines (USVMs) with Székely regularization	Uplift algorithm	Therapy with right heart catheterization procedure (RCH)	Clinical trials	Open-source real-world data[61]
Jaskowski and Jaroszewicz[31]	Response variable transformation	Uplift modeling strategy	Three experiments: (1) Therapy with peripheral blood transplant; (2) Therapy with tamoxifen and radio therapy against breast cancer; (3) Therapy with steroids against hepatitis	Clinical trials	Open-source real-world data[60,62]

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 2. (*Continued*)

Study	Main topic	Research stream	Experiment	Industry/Science	Data origin
Kane *et al.*[21]	Generalized weighting procedure of class probabilities, comparison of uplift approaches; signal-to-noise (S/N) ratio	Uplift modeling strategy	Three experiments: (1) Direct mail (paper); (2) Email; (3) Direct mail (paper)	Financial services, e-merchandise, retail office supplies	Private real-world data (anonymized authority)
Kuusisto *et al.*[51]	USVMs	Uplift algorithm	Simulated marketing activity	—	Simulation data
Lai *et al.*[47]	Transformation scheme with weighted class probabilities	Uplift modeling strategy	Loan product promotion	Financial services	Canadian Imperial Bank of Commerce (CIBC)
Larsen[49]	Uplift k-nearest neighbor and variable selection	Uplift algorithm	—	—	—
Lo[33]	Interaction term approach	Uplift modeling strategy	—	—	Simulation data
Lo and Pachamanova[63]	Multiple treatment optimization approach for prescriptive uplift analytics	Uplift algorithm	Email campaign (men and women separately targeted)	Online retail	Open-source real-world data[58]
Manahan[48]	Uplift neural network implementation with SAS	Uplift algorithm	Contract renewal campaign	Telecommunication	Private real-world data (Cingular)
Nassif *et al.*[64]	Multi-relational uplift modeling system for medical research (SAYL algorithm)	Uplift algorithm	Therapy against breast cancer	Clinical trials	Open-source real-world data,[67] 2012)

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 2. (*Continued*)

Study	Main topic	Research stream	Experiment	Industry/Science	Data origin
Nassif *et al.*[39]	Alternative uplift evaluation measures (ROC)	Specific task	Therapy against breast cancer	Clinical trials	Open-source real-world data[68]
Radcliffe[18]	Uplift evaluation measures	Specific task	Three experiments: (1) Catalogue mailing; (2) Retention marketing; (3) Cross-selling	Retail, telecommunication, financial services	Private real-world data (anonymized authority)
Radcliffe and Surry[30]	Fundamental idea of uplift modeling with reference to differential response analysis	Uplift algorithm	—	—	—
Radcliffe and Surry[42]	Significance-based uplift decision trees with several key features, uplift evaluation measures	Uplift algorithm	—	—	—
Rzepakowski and Jaroszewicz[36]	Uplift modeling for multiple treatments	Uplift algorithm	No campaign conducted (artificial allocation of observations to either treatment or control group in 16 datasets)	—	Open-source real-world data[60]
Rzepakowski and Jaroszewicz[17]	Uplift decision trees with different split criteria	Uplift algorithm	Email campaign (men and women separately targeted)	Online retail	Open-source real-world data[58]
Shaar *et al.*[46]	Pessimistic uplift modeling approach to minimize disturbance effects	Uplift modeling strategy	Simulated campaigns/ treatments in marketing and medicine	—	Open-source real-world data[58,60,62]

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 2. (*Continued*)

Study	Main topic	Research stream	Experiment	Industry/Science	Data origin
Sołtys *et al.*[45]	Ensemble methods for uplift modeling (bagging, random forest)	Uplift algorithm	Simulated campaigns/ treatments in marketing and medicine	—	Open-source real-world data[58,60,62]
Su *et al.*[65]	Causal inference trees and uplift k-nearest neighbor approach in assessing treatment effects	Uplift algorithm	Synthetic data creation (uniform distribution)	Machine learning research	Simulation data
Tian *et al.*[34]	Investigation of the effects of a transformation of input space on a certain outcome of interest in medical research	Uplift modeling strategy	(1) Study of the implications of ACE inhibitors on lowering cardiovascular risk for patients with stable coronary artery disease and normal or reduced left ventricular function (2) Study of interactions between gene expression levels and Tamoxifen treatment in breast cancer patients	Clinical trials	(1) Preventive of Events with Angiotension Converting Enzyme Inhibition (PEACE) study[69] Breast cancer dataset consisting of 414 patients in the cohort GSE6532[70]
Yong[66]	Prediction inference procedure with stratification to obtain generalizable predictions for medical examinations	Specific task	Several	Clinical trials	Several; among them the Mayo liver study data
Zaniewicz and Jaroszewicz[37]	USVMs	Uplift algorithm	Simulated campaigns/ treatments in marketing and medicine	—	Open-source real-world data[58,60,62]

*https://cran.r-project.org/web/packages/Information/index.html; **https://www.udemy.com/uplift-modeling

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

further empirical testing. As described above, tree-based algorithms are especially popular and used, amongst others, by Radcliffe and Surry[30]; Hansotia and Rukstales[40]; Chickering and Heckerman.[41] Other studies consider base learners such as logistic regression,[33] neural networks,[48] and *k*-nearest-neighbors.[49] We also observe some authors to use support vector machines for uplift modeling.[37,50,51] Due to the focus of previous research to consider a specific learning algorithm, empirical evidence related to interactions between uplift modeling strategies and learning algorithms is lacking. Therefore, one objective of this paper is to implement uplift modeling strategies using a set of alternative classification algorithms, which we believe to offer original insights related to the relative suitability of different learners to implement specific uplift modeling strategies.

The overarching conclusion emerging from Table 2 for e-commerce in general and marketing campaign planning is that available approaches in uplift modeling come from diverse strands of literature. This motivates a systematic comparison of the performance of such approaches, which, according to Table 2, is missing. Given that marketers typically use response models,[9] modification of such models to account for uplift effects would mean additional efforts and, eventually, sacrifice of well-timed performance. In contrast to the individual development of uplift algorithms, we therefore regard uplift modeling strategies as more beneficial for e-commerce since they make it possible to apply several supervised learning algorithms for uplift modeling without the need of modification. From current literature, we observe that there are only few papers that focus on such strategies and an empirical comparison of available strategies with several supervised learning algorithms is missing. Instead, studies in uplift modeling focus on single models. This explains the lack of specific recommendations of which models are comparably most valuable to apply. To close this research gap, we benchmark available modeling strategies for conversion uplift that we pair with multiple base learners.

4. Uplift Modeling Strategies

The paper empirically compares eight uplift modeling strategies, which we depict in Table 3. The strategies have been proposed in previous work in different settings. Evidence on their relative effectiveness in a given context is lacking and originally provided here. We reintroduce the modeling strategies in subsequent sections and distinguish between basic, advanced, and special strategies for conversion uplift. The latter exhibits a comparable level of complexity as advanced strategies but does not necessarily focus on data transformation schemes. Rather, related strategies have their own distinct characteristics and are based on most recent research.

With the choice of strategies, we are confident to provide a wide portfolio of state-of-the-art uplift modeling strategies. Recall that the strategies enhance execution of standard classification procedures for uplift modeling. As a result, the strategies can be practiced directly in e-commerce initiatives such as customer acquisition, customer development,[21] or customer retention[35] without a need to modify base learners.

Table 3. Uplift modeling strategies.

Category	Uplift modeling strategy	Acronym	Source
Basic	Two-Model Uplift Method	TWO_MODEL	*Various*
Advanced	Interaction Term Method	ITM	Lo[33]
	Treatment-Covariates Interactions Approach	TCIA	Tian *et al.*[34]
	Class Variable Transformation	CVT	Jaskowski and Jaroszewicz[31]
	Lai's Weighted Uplift Method	LWUM	Lai *et al.*[47]
	Lai's Generalized Weighted Uplift Method	LGWUM	Kane *et al.*[21]
Special	Reflective Uplift Modeling	REFLECTIVE	Shaar *et al.*[46]
	Pessimistic Uplift Modeling	PESSIMISTIC	Shaar *et al.*[46]

Consider a training set $\text{TRAIN}_m = \{(x_i, y_i)\}_{i=1}^{m}$ of m customers gathered, for example, by means of a pilot campaign. Every customer is characterized by a set of explanatory variables x_i and a binary variable $y_i \in \{0, 1\}$ that indicates whether a conversion has been observed. We refer to y_i as the target variable that we seek to explain. Let T_i and C_i indicate the membership of customer i to the treatment or control group, with prior probability distributions $P(T_i)$ and $P(C_i)$. Then, $P(Y_i = 1|T_i X_i)$ and $P(Y_i = 1|C_i X_i)$ denote the conditional probabilities of conversion for treatment and control group customers, respectively. For notational convenience, we refer to these conditional probabilities as $P(Y_i|T_i)$ and $P(Y_i|C_i)$ in the following. Furthermore, we define the four unconditional probabilities as follows: $P(T_i \cap Y_i)$ treated and response, $P(T_i \cap \bar{Y}_i)$ treated and non-response, $P(C_i \cap Y_i)$ non-treated and response, and $P(C_i \cap \bar{Y}_i)$ non-treated and non-response.

4.1. *Basic uplift modeling strategy*

The ***two-model*** uplift method (e.g., Refs. 18 and 30) captures the difference in class probabilities by providing a mechanism to differentiate between structures of customers' motivation:

$$Uplift_i^{TWO\_MODEL} = P(Y_i|T_i) - P(Y_i|C_i). \tag{1}$$

Building and predicting with two equal learning algorithms given these two samples constitutes the methodology of the two-model uplift method. In contrast, response models predict $P(Y_i|T_i)$.

4.2. *Advanced uplift modeling strategies*

Lo[33] proposes a modification of the explanatory variables. He introduces a dummy variable $D_i \in \{0, 1\}$ for control and treatment group, respectively. D_i is multiplied with the entire input space X_i to gain an interaction term that is used in model prediction:

$$Uplift_i^{ITM} = P(Y_i|X_i, D_i, X_i \cdot D_i). \tag{2}$$

More specifically, ***ITM*** of Lo[33] first develops an uplift model from the training data where D_i is known for all customers. Unlike the two-model uplift method, which

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

develops individual classification models for treatment and control group customers, ITM estimates only one model. Then, to estimate uplift for a novel customer with characteristics X_i^{new}, this single model is evaluated twice; setting $D_i = 1$ and $D_i = 0$ in the first and second evaluation, respectively. As is evident from (2), the resulting probability predictions will differ because of D_i. The former represents the customer's conversion probability if treated while the prediction resulting from setting $D_i = 0$ approximates the conversion probability without treatment. Similar to the two-model uplift method, the estimate of conversion uplift is given by the difference between the two predictions.

Independently from Lo,[33] Tian *et al.*[34] propose an uplift modeling strategy, called ***TCIA*** in the following, which is conceptually similar to ITM. Differences to ITM are minute and limited to the coding and scaling of the interaction terms. In particular, Tian *et al.*[34] obtain a set of interaction terms, D_i^*, as $D_i^* = \frac{X_i^* \cdot D_i}{2}$, whereby X_i^* denotes the original covariates, X_i, after mean centering. Another difference relates to D_i, which in the case of ITM, represents a zero-one dummy variable for control and treatment group, respectively. Thus, ITM captures differences in the treatment effect via movements of the intercept. This differs from TCIA where Tian *et al.*[34] set $D_i \in \{-1, 1\}$. As a result, TCIA captures the treatment effect by subtracting treatment and control group probabilities within one functional form. With these modifications, TCIA predicts uplift as:

$$Uplift_i^{TCIA} = P(Y_i | X_i, D_i^*). \tag{3}$$

Tian *et al.*[34] have applied their uplift modeling strategy to study interactions between gene expression levels and drug substances regarding breast cancer patients. Guelman *et al.*[54] further validated this modeling strategy by means of a simulation.

Jaskowski and Jaroszewicz[31] present a transformation procedure — ***CVT*** — that develops a novel target variable based on the original target (i.e., binary conversion response) and the membership of the respective customer to either the treatment or control group. Let Z_i denote the binary transformed target variable corresponding to customer i. Then, $Z_i = 1$ if $(T_i \cap Y_i) \cup (C_i \cap \bar{Y}_i)$ is given; otherwise $Z_i = 0$. Thus, $Z_i = 1$ captures treated customers with response as well as non-treated customers without response. On the contrary, for treated customers without response as well as non-treated customers with response, $Z_i = 0$. The definition of Z_i is based on the link between the desired behavior and a marketing action. More specifically, $Z_i = 1$ reflects customers that convert due to an incentive, but do not convert if not being solicited. The focus of this modeling strategy is to target these customers because they are likely to be persuaded. Hence, as opposed to previous uplift modeling strategies, the uplift effect is based on the distribution of the transformed conversion variable and defined as:

$$Uplift_i^{CVT} = 2 \cdot P(Z_i = 1) - 1. \tag{4}$$

Lai *et al.*[47] presents an extension of CVT — ***LWUM*** — that weights probabilities of positive and negative classes. LWUM assumes that the positive uplift lies in

correctly identified *persuadables* (here, treatment-group responders and control-group non-responders), whilst the negative uplift can be found in the *do-not-disturbs* group (here, treatment-group non-responders and control-group responders). Therefore, let W be the number of positive observations divided by the total population. The uplift effect is then defined as:

$$Uplift_i^{LWUM} = P(Z_i = 1) \cdot W - P(Z_i = 0) \cdot (1 - W). \tag{5}$$

LWUM, thus, seeks to maximize the positive uplift while decreasing negative uplift in the first decile.

Kane *et al.*[21] present ***LGWUM*** as the generalized version of LWUM with weighted probability scores that realize the influence of the fraction of treatment and control group customers on the lift measure and is defined as:

$$Uplift_i^{LGWUM} = P(Y_i|T_i) + P(\bar{Y}_i|C_i) - P(\bar{Y}_i|T_i) - P(Y_i|C_i). \tag{6}$$

4.3. *Special uplift modeling strategies*

Shaar *et al.*[46] present the ***reflective*** uplift modeling strategy by two separate models that are built to learn the treatment effect in the conversion and non-conversion groups. The authors recognize disturbance effects when applying uplift models. The first one is a response effect that takes place due to correlation between explanatory variables and a binary class label, and the second effect — a partitioning effect — appears when the treatment indicator depends on the covariates. To overcome these negative effects, reflective uplift modeling has been introduced. The uplift effect is then calculated, whereas the groups are treated as positive and negative as in CVT:

$$Uplift_i^{REFLECTIVE} = P(T_i \cap Y_i) \cup P(C_i \cap \bar{Y}_i) - P(T_i \cap \bar{Y}_i) \cup P(C_i \cap Y_i). \tag{7}$$

Thus, the probabilities for positive and negative groups are obtained from two different models. To determine a score in terms of pessimistic uplift modeling, LWUM is again considered. The final ***pessimistic*** uplift modeling strategy is defined as:

$$Uplift_i^{PESSIMISTIC} = 0.5 \cdot (Uplift_i^{LWUM} + Uplift_i^{REFLECTIVE}). \tag{8}$$

4.4. *Conceptual evaluation*

In this section, we examine the relative merits of the modeling strategies for conversion uplift from a conceptual perspective. First, we consider the two-model uplift method that is presented by the difference between the class probabilities (i.e., treatment versus non-treatment). This modeling strategy suffers from poor approximation, since both probability estimates originate from two separate samples (e.g., Refs. 20 and 53). ITM[33] and TCIA[34] manipulate the data input space through interaction terms with dummy variables indicating the treatment effect.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Incorporating interaction effects for all variables, these uplift modeling strategies increase dimensionality. Therefore, ITM and TCIA appear less suitable for datasets where the number of original variables is large. CVT as in Jaskowski and Jaroszewicz,[31] on the contrary, changes the response variable to facilitate focusing on *persuadables* and *do-not-disturbs* and improves targeting decisions. However, CVT does not regard the difference between the relative sizes of positive and negative observations. This is why Lai *et al.*[47] introduced the weights as per proportion of positive and negative observations and developed LWUM. This uplift modeling strategy should address differences in these proportions. However, we expect the accuracy of LWUM to suffer when the ratio of treatment and non-treatment assignments is not approximately equal. LGWUM[21] overcomes this and is designed to combat disturbance such as multicollinearity. Shaar *et al.*[46] presents the reflective uplift modeling strategy that estimates the uplift effect from the conversion and non-conversion groups. The authors further extend it through pessimistic uplift modeling that combines LWUM and the reflective uplift modeling strategy into one model. The combination is claimed to overcome the disadvantage of the two-model uplift method, where the separated estimation of response probabilities among treatment and control group customers deteriorates the accuracy of uplift predictions (e.g., Refs. 20, 46 and 53).

5. Experimental Setup

We involve numerous datasets that belong to the field of e-commerce, indicating the goal to categorize the customer base into two classes: buyer and non-buyer. In the following, we elaborate the campaign process and underlying data, base learners to be paired with the aforementioned uplift modeling strategies, and finally the performance metrics.

5.1. *Campaign process and data*

The experimental setup involves 27 datasets from several digital marketing campaigns. These campaigns were executed by Akanoo,[b] a company specializing in analytics-as-a-service solutions for online shops. Akanoo provided us with a fully anonymized version of real-world campaign data in the scope of a research collaboration. The data is sensitive and can, therefore, not be disclosed to the public. It includes multiple campaigns that were carried out in different electronic marketplaces and designed so that customers showing specific behavioral patterns during their shop visit, as identified by an uplift model, are targeted with a digital coupon. Customers that leave the respective shop by having activated this coupon obtain a discount of 10% off their final basket value. A real-time targeting process has been applied to identify customers to receive the coupon. Every customer has been assigned either to the treatment or control group by chance or by a model. In the

[b]https://akanoo.com/.

Table 4. Summary of e-retail datasets.

Shop	Product line	Geographical location	No. of cases: treatment/control	No. of responses: treatment/control	Uplift (%)
1	Apparel	Poland	206,148/69,177	6,909/2,289	0.04
2	Apparel	Germany	128,469/43,467	8,277/2,523	0.64
3	Apparel	Germany	36,288/12,327	3054/879	1.29
4	DIY products	United Kingdom	216,534/72,978	5,160/1,560	0.25
5	Apparel	Czech Republic	46,983/16,284	2,733/1,005	−0.35
6	Apparel	Germany	8,733/2,877	786/234	0.87
7	Books and multimedia	Germany	9,003/3,030	360/114	0.24
8	Toys	Germany	898,734/300,847	96,318/31,874	0.12
9	DIY products	Germany	92,961/31,125	1,800/525	0.25
10	DIY products	France	9,471/3,309	501/129	1.39
11	Pharmaceuticals	Germany	5,319/1,680	2,436/807	−2.24
12	Special apparel (hats)	Germany	16,734/5,580	1,911/603	0.61
13	Apparel/household items	France	47,964/15,900	135/24	0.13
14	Fan articles and toys	Germany	9,534/3,303	777/168	3.06
15	Apparel	Germany	18,417/6033	2,472/708	1.69
16	Apparel	The Netherlands	5,520/1,806	348/75	2.15
17	Alcoholic beverages	Germany	6,996/2,400	1,803/624	−0.23
18	Pharmaceuticals	Germany	6,699/1,998	3,411/990	1.37
19	Sports apparel/accessories	Germany	83,865/27,765	13,428/4599	−0.55
20	Pet food	Germany	16,881/5,601	3,456/1,143	0.07
21	Apparel	Germany	89,424/30,141	6,060/1,926	0.39
22	Shoes and accessories	Germany	244,506/81,726	14,643/5,031	−0.17
23	Pharmaceuticals	Germany	4,104/1,239	2,304/651	3.60
24	Apparel	Austria	20,913/6,855	684/207	0.25
25	Shoes	Germany	2,403/801	99/39	−0.75
26	Special apparel (hats)	The Netherlands	7,863/2,589	396/114	0.63
27	Outdoor apparel	Germany	45,210/14,928	2,469/846	−0.21

latter case, the individual online behavior of new customers is considered after five pageviews and that of returning customers after three pageviews. The derived predictive scores determine whether the customer is likely to be persuadable (i.e., customer with high probability to respond if being treated with coupon). As a result, the model qualifies the customers to the treatment group. The systematic component of the targeting process creates a selection bias that leads to a quasi-experiment.

Table 4 summarizes the available datasets in terms of product line, geographical location, the number of observations, and responses in the treatment and control group, respectively, and the uplift.

Table 4 indicates that the consumer goods relate to different sorts of apparel, toys, garden articles, books and multimedia, pet food, and many other. In addition, sports and outdoor articles are also sold in a few shops. Businesses operate in Austria, the Czech Republic, France, Germany, The Netherlands, Poland, and the United Kingdom. The total number of cases across all datasets is roughly 3 million. On an average, we observe a treatment to control group ratio of 3:1 meaning that three out of four customers have received a digital coupon. Based on the number of responses in the treatment and control group, we capture the impact of the shop-wise

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

marketing campaigns. The last column in Table 4 reports the campaign uplift per shop, which we calculate as the differences between the relative response rate in the treatment and control group, respectively. Table 4 reveals low (positive) uplift for almost every shop. For some shops, we even observe negative uplift resulting from the response rate in the control group exceeding the response rate in the treatment group. The average uplift across the 27 shops is 0.54%.

The data contain 60 features that profile the customers' behavior. Every observation relates to the shop-based journey performed during a certain time span (i.e., from entering to leaving the shop). Cookie technology allows to differentiate between new and returning customers. Most features are numeric while some are factors. These features provide information on numerous customer activities during the shop visit, for instance, how much time the customer spends on certain page types, whether the customer has purchased a specific product in the same shop in the past and how much time has passed since the customer added an item to the shopping cart. Further examples relate to how many views the customer has made on a sale-related page and how many products lie in the customer's shopping basket during the current session. Inspiration on data collection has been gained from Van den Poel and Buckinx.[71] Furthermore, meta dimensions that are crucial for the use of uplift models have been collected, particularly, an indicator of treatment or control group assignment, and a variable that captures the purchase event. Being structured in terms of current session, previous session(s), and identifiables of the respective customer, Table 5 lists and describes all features used for the empirical study.

Another important concern relates to data partitioning. We have created three partitions from the available data: 40% training partition that we use to train the strategies, 30% for a parameter-tuning partition that we use to validate the meta-parameter tuning, and another 30% for a test partition. To guarantee a reliable evaluation, we apply a 10-fold cross validation scheme "through time" to reflect the situation in marketing practice and increase the size of observations by resampling. For all uplift modeling strategies, the stated models first predict the training and parameter-tuning partitions together. Strategy-wise models with the best candidate settings are then validated on the validation sample to assure a reliable benchmark.

5.2. *Base learners*

The experimental design includes six base learners to ensure a vast benchmark study. Recall that we benchmark modeling strategies for conversion uplift that can be paired with any base learner. Thus, we secure every possible combination between uplift modeling strategy and base learners. The experiment is performed in a Python environment and builds upon popular libraries for data manipulation, statistics, visualization, and data science; namely NumPy, Pandas, Matplotlib, and Scikit-learn.[72] To guarantee robust results, we execute a wide range of meta-parameters for every base learner, presented in Table 6. Every model is tuned automatically and

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 5. Clickstream features used for empirical study.

Setting	#	Feature name	Description	Based on Van den Poel and Buckinx[71]
Current session	45	InitBasketNonEmpty	State of the initial basket (empty vs. non-empty)	
		HadBasketAdd	Whether the visitor has added at least one product to the basket	
		TimeToBasketAdd	Amount of time since a product has been added to the basket	
		BasketQuantity	Number of products in current basket	
		NormalizedBasketSum	Normalized value of customer basket (for comparisons across shops)	
		TimeToFirst (pagetype)	Time span from shop arrival to first click on page type "cart"/"overview"/"product"/"sale"/"search"	
		TimeSinceFirst (pagetype)	Amount of time since first click on page type "cart"/"overview"/"product"/"sale"/"search"	
		TimeSinceOn (pagetype)	Duration on page type "cart"/"overview"/"product"/"sale"/"search" until leave of online shop	
		TimeOn (pagetype)	Duration on page type "cart"/"overview"/"product"/"sale"/"search" until leave of page type	
		HourOfDay	Hour of the day (1–24) when the visitor has entered the online shop	X
		SessionTime	Duration of current visitor session	
		ScrollHeight (overview)	Scroll height for pages of type "overview"	
		ScreenWidth	Screen width of customer device	
		TabSwitch (product)	Number of total tab switches for pages of type "product" during session	
		Clicks (product)	Number of clicks for pages of type "product"	X*
		TimeSinceClick	Time span from first click to shop leave	
		TimeSinceTabSwitch	Time span from first switch of tabs	
		ViewCount	Number of views in the current session	X*
		ViewedBefore (cart)	Whether the visitor has already viewed a specific page from page type "cart"	
		ViewsOn (pagetype)	Number of views on page type "cart"/"overview"/"product"/"sale"/"search"	X*
		InitPageWas (overview)	Whether the initial page had page type "overview"	
		InitPageWas (product)	Whether the initial page had page type "product"	
		InitPageWas (sale)	Whether the initial page had page type "sale"	
		NumberOfDifferent Pages (overview)	Number of views on different pages from page type "overview"	X*
		NumberOfDifferent Pages (product)	Number of views on different pages from page type "product"	X*

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 5. (*Continued*)

Setting	#	Feature name	Description	Based on Van den Poel and Buckinx[71]
Previous session(s)	8	TimeSinceLastConversion	Amount of time since last product purchase	X
		VisitCountLastWeek	Number of shop visits within the previous week	
		VisitCountToday	Number of shop visits during the day of current session	
		PreviousVisitCount	Number of previous shop visits	X
		TimeSinceFirstVisit	Amount of time since first shop visit	
		TimeSinceLastVisit	Amount of time since last shop visit	X*
		DurationLastVisit	Time span of previous shop visit	X*
		ViewCountLastVisit	Number of views during last shop visit	X*
Identifiables	7	VisitorKnown	Whether the visitor has already visited the shop in the past	X*
		WasConvertedBefore	Whether the visitor has already purchased a product in a previous session	
		Conversion	Whether the visitor has purchased a product in the current session	
		Normalized revenue	Normalized amount of revenue (for comparisons across shops)	
		Treatment/control group	Whether the visitor has received an e-coupon	
		Shop-ID	Unique shop identifier	
		Timestamp	Point in time when visitor has entered the online shop	

*Based on Van den Poel and Buckinx[71] but slightly adapted.

Table 6. Meta-parameters of the base learners.

Base learner	Acronym	No. of models	Meta-parameter	Candidate setting
Logistic regression	LogR	34	Regularization term Regularization factor	[L1, L2] [1e-8, 1e-7, . . ., 1e8]
Support vector machines with linear kernel	SVC	42	Regularization factor Calibration method	[1e-10, 1e-9, . . ., 1e10] [Sigmoid, Isotonic]
k-nearest neighbor	KNN	20	Number of nearest neighbors	[1, 5, 10, 20, . . ., 100, 200, . . ., 500, 1000, 2000, . . ., 4000]
Naïve Bayes	NB	1	—	—
Stochastic gradient descent for classification	SGDC	144	Loss function Regularization term Alpha Learning rate	[Log, Mod. Huber, Hidge, Percep.] [L1, L2, Elastic Net] [1e-6, 1e-5, . . ., 1e-1] [Optimal, Invscaling]
Random forest for classification	RFC	4	Max. no. of covariates Min. no. of samples	[8, 9] [1000, 2000]

transmitted to the cross-validation technique discussed previously. In total, we involve 245 models.

We pair base learners and modeling strategies for conversion uplift in a full-factorial experimental setup. Recall that we consider eight uplift modeling strategies and response modeling. In total, we produce (8+1) modeling strategies * 245 base learners = 2,205 models. We choose the base learners due to their popularity in response and uplift modeling. In response modeling, for example, they are often questioned in pivotal benchmark studies.[73,74] SGDC and RFC demonstrate excellent performance in real-world experiments.[35] Due to the fact that RFC is less sensitive to meta-parameter adaptations than SGDC,[75] we consider for RFC a smaller number of models. In uplift modeling, LogR,[33] KNN,[49] and SVC[37,51] have gained a strong research interest. As a standard base learner without meta-parameters, we add a NB algorithm to the library of base learners.

5.3. *Validation measures*

Typically, the performance of predictive models grounds on a comparison of actual versus predicted outcomes. In uplift modeling, however, this is not reasonable since a customer cannot be part of both the treatment and control group. This phenomenon is known as the fundamental problem of causal inference.[76] Consequently, today's best practice is a decile-based evaluation approach to identify uplift. Hence, model performance is captured in terms of Qini coefficient Q and visualized in uplift gains charts by means of Qini curves.[18] This includes the assumption that similarly scored cases behave likewise, i.e., the k percent highest scores on treatment out-of-sample test data are compared to the k percent highest scores on control out-of-sample test

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

data and with the subtraction of the top gains from both groups a meaningful estimate of uplift can be derived.[31] Q is, thus, defined as the area between a model's Qini curve and a random targeting line.[42] Because typically uplift gains charts display Qini curves that relate to a cumulative measure, we further consider uplift bar charts that mask the effect of cumulativeness to provide a decile-isolated analysis of model performance.

6. Empirical Results

The experimental results consist of the performance estimates for every combination of 6 levels of base learners, 9 levels of modeling strategies (response modeling included), and 27 levels of datasets. The performance measures capture the degree to which the marketing campaign strategy improves via application of uplift modeling strategies in terms of Qini coefficient and cumulative (non-cumulative) number of incremental purchases.

6.1. *Examination of the interaction between uplift modeling strategies and base learners*

To identify synergy effects between the modeling strategies for conversion uplift and base learners, we examine their interaction in Table 7. In particular, we pair every base learner with all uplift modeling strategies and capture the predictive performance on the out-of-sample test set in terms of Qini coefficient. These values are averaged over the datasets. We express the Qini coefficient in percentage terms, i.e., Q_{pct}, by subtracting the control group response rate from the treatment group response rate for every decile. In contrast to the general Q coefficient,[42] Q_{pct} makes comparisons across the datasets with different number of observations possible and, thus, requires no normalization procedure. To increase the readability of Q_{pct}, we multiply its values with a factor of 1,000. We use bold face for every best combination (i.e., uplift modeling strategy coupled with base learner). For example, the value in the last column for CVT is marked in bold face indicating that CVT interacts best with RFC.

Table 7. Qini coefficient of uplift modeling strategies.

Uplift modeling strategy	Base learner					
	KNN	LogR	NB	SGDC	SVC	RFC
CVT	3.171	3.348	−0.951	−1.041	2.017	**6.145**
ITM	3.991	2.901	3.770	0.979	**8.017**	3.216
LGWUM	−0.230	3.767	−4.459	1.831	−0.932	**5.593**
LWUM	3.171	4.258	−0.945	0.203	2.049	**6.130**
PESSIMISTIC	1.418	4.269	−1.626	0.720	2.010	**6.606**
REFLECTIVE	−1.526	**3.310**	−2.914	0.868	−0.727	2.303
TCIA	1.043	−1.950	−2.821	1.222	**3.893**	0.403
TWO_MODEL	**7.267**	4.305	3.297	0.688	2.806	5.401

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 7 reveals multiple important findings. First, the best possible interaction is between ITM and SVC with Q_{pct} of 8.017. This is followed by the two-model uplift method coupled with KNN with Q_{pct} of 7.267 and CVT with RFC of 6.145. This strongly signals in favor of ITM as a modeling strategy for conversion uplift and of SVC as a base learner. This view is only strengthened when we look at the pair of TCIA and SVC, where SVC is the best performer. However, we recommend RFC as a base learner for uplift modeling since it collects the biggest number of wins. More specifically, RFC is the best performer when coupled with CVT, LGWUM, LWUM, and pessimistic uplift modeling. We observe that KNN performs best when paired with the two-model uplift method and the differences in the performance compared to other uplift modeling strategies are substantial. For example, the pair of the two-model uplift method and KNN achieves Q_{pct} of 7.267 compared to the second-best performing pair of ITM and KNN with Q_{pct} of 3.991 and the worst performer pair of the reflective uplift modeling strategy and KNN with Q_{pct} of -1.526. As a result, we can only recommend considering KNN when coupled with the two-model uplift method. We also observe that the reflective uplift modeling strategy performs best coupled with LogR. However, LogR shows also high and better potential when interacting with other strategies. For example, Q_{pct} of couples of pessimistic uplift modeling, LWUM, and the two-model uplift method with LogR is higher than that of reflective uplift modeling with LogR. Thus, LogR seems to be more flexible than KNN for uplift modeling. On the contrary, due to the weak performance compared to other base learners, NB and SGDC have no wins. Thus, we cannot recommend executing them for uplift modeling. This recommendation is supported by the fact that for many uplift modeling strategies, NB collects negative Q_{pct} values. The same applies to the pair of CVT and SGDC. We also would like to stress that the best pessimistic uplift model outperforms all base learners related to LWUM and reflective uplift modeling. This is interesting since LWUM and reflective uplift modeling hold equal shares in creation of the pessimistic modeling strategy. LGWUM does not add more value than LWUM. With SGDC being the only exception, all base learners paired with LWUM outperform their equivalents for LGWUM. A similar picture emerges for covariate transformations. All base learners but SGDC and SVC paired with CVT obtain higher Q_{pct} values when compared to respective TCIA counterparties.

To further support findings obtained from Table 7, we compare the performance of the modeling strategies through a robustness procedure. In particular, we capture the performance of the uplift modeling strategies coupled with base learners in a 10-fold cross validation and visualize it in Fig. 1. Every boxplot portrays base learners on the x-axis and the performance measured in Q_{pct} on the y-axis. We scale the Q_{pct} values to ease comparability.

Figure 1 supports some previous findings but also reveals new ones. First, we would like to highlight the remarkable performance of RFC. RFC is the best performer when coupled with, e.g., CVT, LGWUM, or pessimistic uplift modeling. Furthermore, RFC shows relatively small variance. This can be emphasized through

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

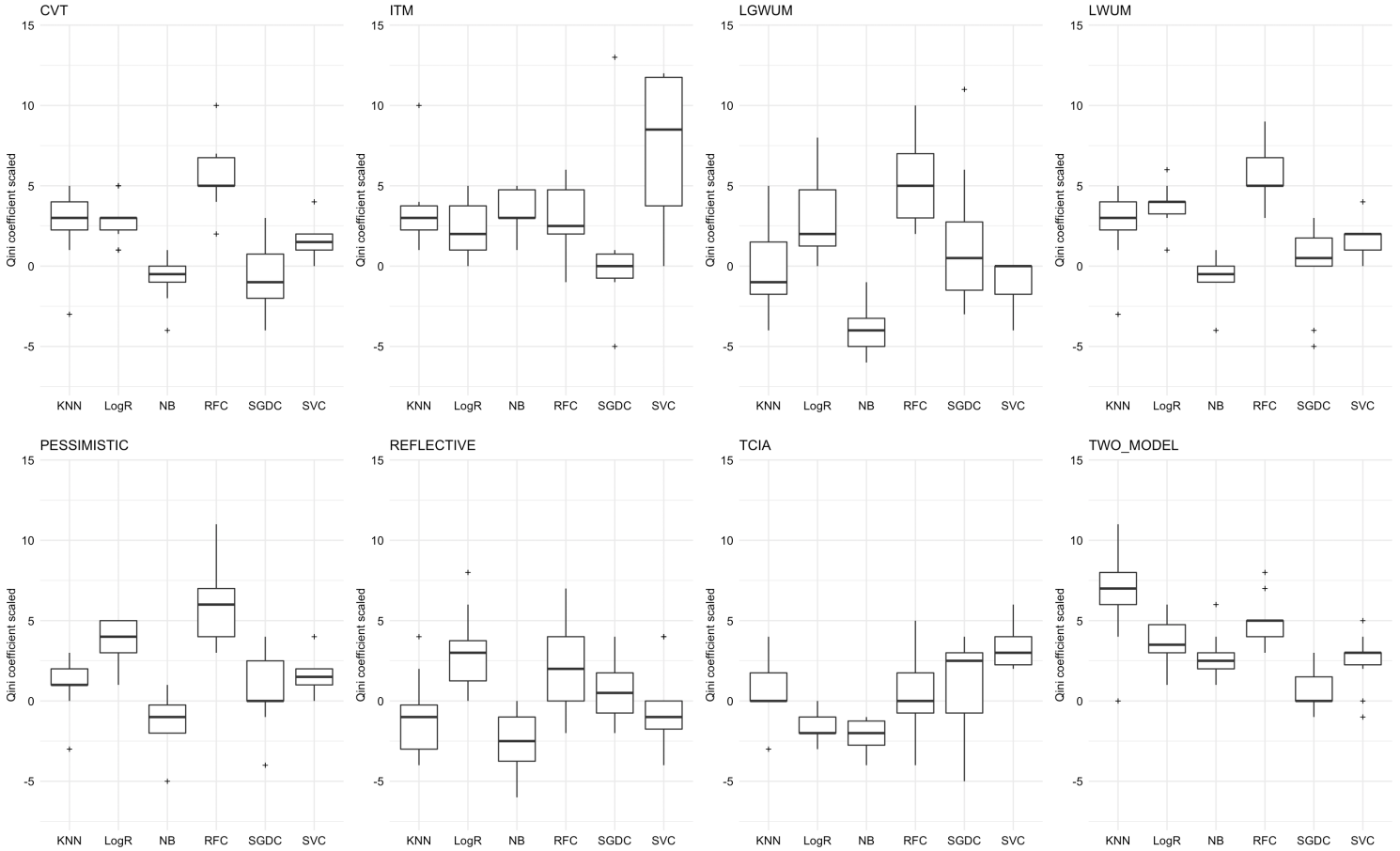

Fig. 1. Scaled Qini coefficient across uplift modeling strategies.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

the combination of RFC with the two-model uplift method. Thus, Fig. 1 further supports the view that RFC is a very suitable base learner for uplift modeling. Second, weak performance of NB and SGDC is further pronounced. We observe that the mean values of NB are negative for the pessimistic and the reflective uplift modeling strategies as well as for CVT. We observe the same pattern on the pair of CVT and SGDC, whereby SGDC also exhibits higher variance than NB. On several uplift modeling strategies apart from CVT, SGDC introduces a comparably high level of variance (e.g., LGWUM, LWUM, TCIA). Whereas gradient descent steps require extensive computational resources in case of large data volume, stochastic gradient descent classifiers are relatively fast to train at the expense of a slower gradient convergence rate, which causes variance from a perspective of single model evaluation. Based on the findings, we caution against using these base learners for uplift modeling. Third, we see a comparably high variance of SVC when coupled with ITM. Due to the high regularization parameter values of several SVCs in our experiment, the SVCs attempt to maximize correctly classified observations at the expense of being more prone towards overfitting. With the goal to reduce predictive bias, consequently, the higher parameter values increase variance (bias/variance trade-off). The variance increase is further encouraged by the special interaction term of ITM, which creates further data points dependent on treatment observations for model estimation. This finding casts doubt on the previous insight where the couple ITM and SVC is the best performer (Q_{pct} of 8.017). Hence, we conclude that ITM paired with SVC does not provide a reliable estimate. In contrast, we observe that KNN paired with the two-model uplift method shows low variance that makes this couple more promising than ITM with SVC (given findings from Table 7). More precisely, ITM-based SVC shows a standard deviation of 8.3 compared to two model-based KNN with 6.0. As a result, KNN has a 28% lower standard deviation than SVC. Note that the standard deviation values are percentages derived from taking the mean of all decile-wise values. At the same time, we conclude that KNN and SVC (ITM being exception) show stable results in terms of variance when coupled with other uplift modeling strategies. The same conclusion can be drawn for LogR which moreover enjoys comparably high stable results across the uplift modeling strategies. In general, we would like to conclude that ITM and the two-model uplift method show the most promising results when interacting with all base learners (SGDC being exception). These uplift modeling strategies do not show negative Q_{pct} values, relatively low variance and comparable results among the base learners. Reflective uplift modeling and TCIA demonstrate opposite performance and, thus, can be regarded as least beneficial modeling strategies involved in this study.

6.2. *Examination of the impact of uplift modeling strategies on business value*

We now examine the potential of the modeling strategies for conversion uplift to increase business value. To do so, we analyze the weighted model performance for

every targeting decile in terms of the cumulative (and later non-cumulative) number of incremental purchases. Again, we describe the effect of every uplift modeling strategy coupled with all base learners. Since the increased number of the incremental purchases results in increased revenue, we argue that uplift modeling might contribute to the increase of business value. To quantify the impact of the modeling strategies on business value, we first provide a tabular view of the decile-wise model performance. Table 8 presents the results obtained on the out-of-sample test set, across the uplift modeling strategies and base learners. For every decile, we highlight the winner among the base learners within the uplift modeling strategy and the global winner (i.e., across all uplift modeling strategies) in italic and bold

Table 8. Summary of cumulative number of incremental purchases.

Uplift modeling strategy/base learner	Cumulative number of incremental purchases per decile									
	1	2	3	4	5	6	7	8	9	10
CVT										
KNN	301	655	831	*1,148*	1,213	1,303	1,332	1,423	1,486	1,671
LogR	819	611	712	795	1,133	1,352	1,401	1,444	1,547	1,671
NB	−234	278	421	213	874	1,098	1,281	1,405	1,533	1,671
SGDC	66	225	422	602	754	918	1,079	1,300	1,440	1,671
SVC	−209	239	484	823	1,158	*1,601*	*1,603*	1,631	1,571	1,671
RFC	*883*	*983*	*1,066*	1,110	*1,297*	1,456	1,597	*1,641*	*1,698*	1,671
ITM										
KNN	418	673	868	901	1,161	1,381	1,395	1,722	1,735	1,671
LogR	*487*	765	1,195	1,332	1,409	*1,755*	1,819	1,690	1,711	1,671
NB	457	687	958	*1,564*	1,080	1,058	1,248	1,445	1,607	1,671
SGDC	96	215	610	741	880	894	783	1,751	***2,223***	1,671
SVC	395	*1,108*	***1,292***	1383	*1,602*	1679	***1,885***	*1,834*	1838	1,671
RFC	482	807	994	1,043	1,198	1,205	1,384	1,341	1,270	1,671
LGWUM										
KNN	347	431	256	499	667	968	1,146	1,392	1,655	1,671
LogR	*591*	*867*	*878*	855	1,127	1,207	1,289	1,564	*1,725*	1,671
NB	364	373	285	11	16	728	901	746	1,039	1,671
SGDC	*434*	642	871	1,002	1,109	1,139	1,226	1,407	1,425	1,671
SVC	271	229	301	609	720	1,010	1,030	1,280	1,432	1,671
RFC	302	689	842	*1,148*	*1,501*	*1,672*	*1,701*	*1,787*	1,713	1,671
LWUM										
KNN	301	655	831	*1,148*	1,213	1,303	1,332	1,423	1,486	1,671
LogR	855	951	909	998	1,143	1,180	1,422	1,436	1,543	1,671
NB	−243	286	408	232	871	1,097	1,282	1,407	1,533	1,671
SGDC	172	385	557	734	869	936	1,150	1,346	1,511	1,671
SVC	−208	245	474	808	1,188	*1,602*	1,617	1,625	1,571	1,671
RFC	*884*	*991*	*1,071*	1,103	*1,294*	1,448	*1,598*	*1,636*	*1,694*	1,671
PESSIMISTIC										
KNN	224	461	719	960	1,065	1,127	1,261	1,357	1,319	1,671
LogR	***935***	932	1,023	934	1,015	1,121	1,350	1,547	1,587	1,671
NB	−148	226	325	674	867	1,005	1,110	1,219	1,127	1,671
SGDC	216	425	655	788	967	1,070	1,124	1,330	1,438	1,671
SVC	−149	226	486	787	1,165	*1,586*	*1,626*	1,577	1,590	1,671
RFC	876	*1,017*	*1,122*	*1,271*	*1,418*	1,495	1,581	*1,581*	*1,685*	1,671

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Table 8. (*Continued*)

Uplift modeling strategy/base learner	Cumulative number of incremental purchases per decile									
	1	2	3	4	5	6	7	8	9	10
REFLECTIVE										
KNN	−123	−6	493	550	813	932	1,140	1,274	1,399	1,671
LogR	*403*	*821*	*970*	*1,022*	1,055	1,236	1,300	1,398	1,584	1,671
NB	50	7	195	192	680	997	1,273	998	1,131	1,671
SGDC	222	378	605	772	975	1,076	1,201	1,374	1,511	1,671
SVC	170	257	368	444	490	816	1,179	1,462	*1,837*	1,671
RFC	−55	276	667	897	*1,150*	*1,464*	*1,544*	*1,500*	1,658	1,671
TCIA										
KNN	133	441	654	795	994	1,220	1,322	1,305	1,371	1,671
LogR	103	60	88	−96	399	962	1,171	*1,573*	*1,922*	1,671
NB	11	229	84	64	463	710	965	1,221	1,838	1,671
SGDC	*309*	*470*	711	843	980	1,065	1,235	1,388	1,482	1,671
SVC	−17	261	*1,033*	*1,190*	*1,281*	*1,677*	*1,578*	1,498	1,685	1,671
RFC	249	423	642	802	961	1,026	1,085	1,157	1,454	1,671
TWO_MODEL										
KNN	321	732	*1,202*	***1,576***	***1,730***	***1,775***	*1,810*	1,760	1,594	1,671
LogR	864	***1,126***	796	1,002	1,257	1,059	1,319	1,503	1,542	1,671
NB	82	583	976	1,183	1,345	1,364	1,410	1,347	1,488	1,671
SGDC	162	421	609	727	929	1,010	1,223	1,378	1,536	1,671
SVC	−49	62	250	900	1,285	1,655	1,785	***1,881***	*1,675*	1,671
RFC	*877*	1,111	1,097	1,117	1,165	1,273	1,437	1,502	1,641	1,671

face, respectively. Consider the very left (upper) column. We contact a 10% fraction of the customer base via marketing campaign. CVT enhances RFC to achieve 883 incremental purchases. We mark this estimate in italic face indicating that RFC is the winner within the 10% fraction across the classifiers paired with CVT. Another example (same column) is the pair of pessimistic uplift modeling and LogR. This pair achieves 935 purchases within the first decile and is marked in both italic and bold face. The former indicates that LogR is the winner regarding the classifiers paired with the pessimistic modeling strategy for the first decile while the latter highlights that the pair of pessimistic uplift modeling and LogR presents the global winner in the first decile across all uplift modeling strategies.

Multiple important findings can be derived from Table 8. First, we would like to emphasize the performance of RFC another time. In particular, we observe that RFC performs well with multiple uplift modeling strategies. For example, within CVT, RFC gets the largest number of wins across the deciles in terms of the cumulative number of purchases compared to the remaining base learners. The same conclusion can be drawn, e.g., for LGWUM and LWUM. RFC is especially successful in the first deciles. Given this, we recommend RFC for the suggestion of Ref. 33 to limit the targeting to the top 10% most valuable customers. However, the success of RFC can be interrupted in the middle deciles. For example, for the 4th decile, the pair CVT and KNN compared to CVT and RFC gets 1,148 and 1,110 cumulative number of

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

purchases, respectively. The pair CVT and SVC outperforms CVT-based RFC in the 6th and 7th deciles. Identical picture can be seen in terms of the pessimistic uplift modeling strategy, whereby SVC gets 1,586 and 1,626 cumulative number of purchases compared to 1,495 and 1,581 of RFC in the 6th and 7th deciles. Thus, we conclude that there are differences in the impact on business value depending on the size of the targeted fraction of the customer base. In general, we see the larger cumulative numbers of purchases in the middle deciles than in the first ones. To give an example, note a steady increase of cumulative purchases for the pair LWUM and SGDC from the first to the last decile. However, this does not necessarily indicate that targeting a larger fraction results in a higher cumulative number of purchases; the pair of the two-model uplift method and KNN in the 7th and 8th deciles (1,810 and 1,760 purchases, respectively) being an example. Therefore, our results clearly show that targeting the whole population of customers — a mail-to-all strategy according to Ref. 41 — is not the best choice. All global winning pairs from the 5th to the 9th decile yield a higher cumulative number of incremental purchases compared to targeting the whole customer base. For instance, the pair of the two-model uplift method and KNN achieves 1,730 incremental purchases by targeting only half of the population while targeting every customer only yields 1,671 incremental purchases.

Most importantly, we now are confident to identify the best combination of base learner and uplift modeling strategy in terms of business value. These pairs are CVT and RFC, ITM and SVC, LGWUM and RFC, LWUM and RFC, pessimistic uplift modeling and RFC, reflective uplift modeling and LogR, TCIA and SVC, and finally the two-model uplift method and KNN. They demonstrate the largest numbers of wins on the deciles. This finding is also supported in terms of Qini coefficient (see 6.1). In the following, therefore, we concentrate on these pairs.

To provide specific recommendations which pair works best, we now present uplift gain charts in Fig. 2. These charts much resemble common gain charts. However, while the performance of models in gain charts in customer acquisition campaigns is typically illustrated by the number of purchases on the y-axis, uplift gain charts draft Qini curves that are by nature capable to signal incrementality. This implies that the number of purchases is replaced by the incremental number of purchases in uplift gains charts. The incremental number of purchases is a helpful indicator to support decision making in marketing practice and can be derived by comparing the purchase rate in the treatment group with the purchase rate in the control group. In both the traditional and uplift case, the purchase indicator is a function of the fraction of people targeted from the campaign's total population, being mapped on the x-axis.[18] Qini curves summarize the decile-wise performance of their underlying uplift models. A diagonal line reflects random targeting and therefore presents a baseline for all uplift modeling strategies. Recall that we present the uplift gain charts only for the winner pairs identified before. We also draw the average performance line — AVG — across the winner pairs to better judge on the performance.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

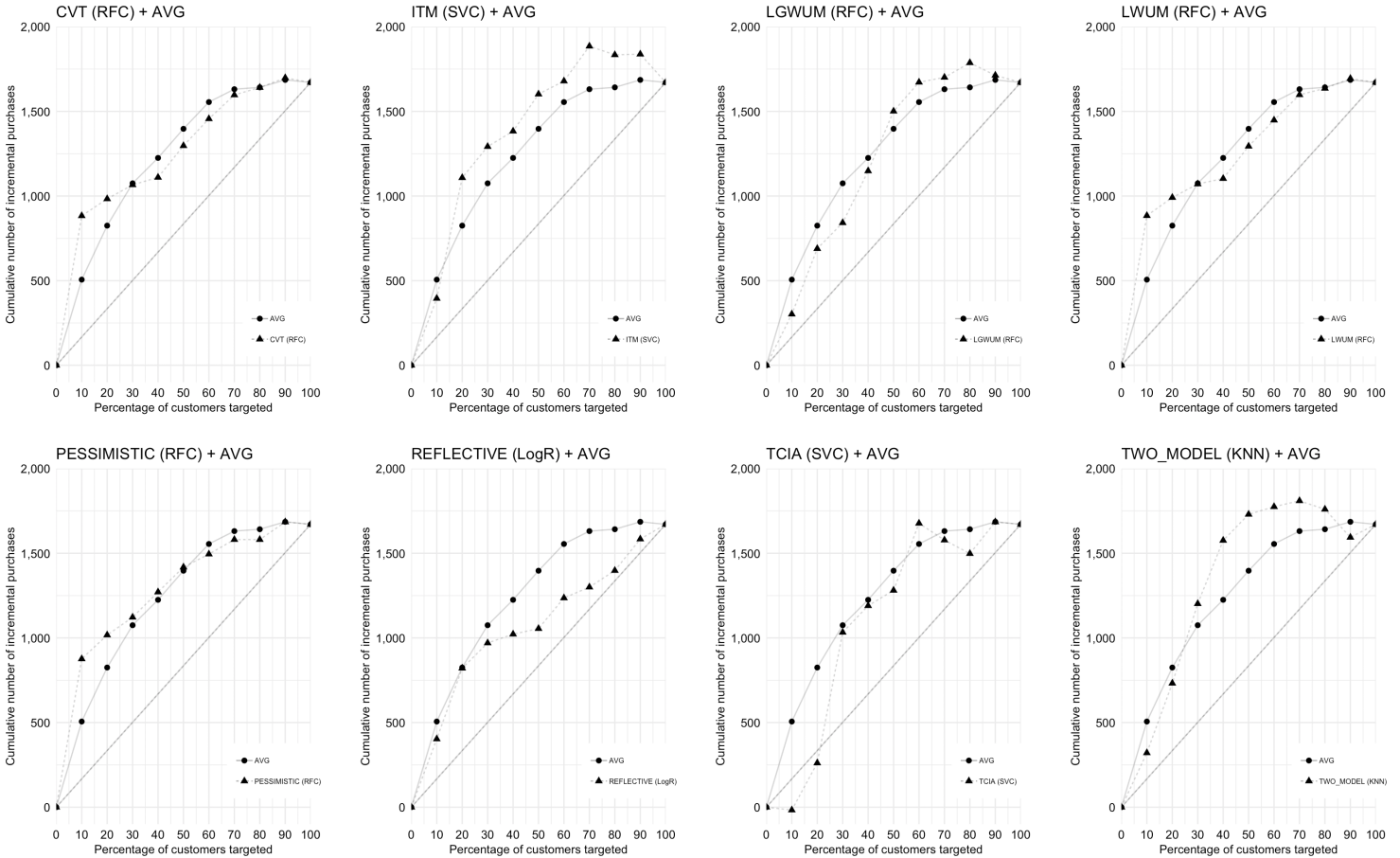

Fig. 2. Uplift gain charts across uplift modeling strategies.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Figure 2 provides new insights into the performance of the winner modeling pairs identified before. First, we observe that all pairs, even though unequally, contribute to higher cumulative number of purchases than the baseline. We see that the higher the fraction of customers targeted, the higher is the cumulative number of purchases in general. However, for several uplift modeling strategies, this is not applicable to the targeting of the whole customer base as targeting a dedicated subgroup of customers yields a higher incremental number of purchases. The limited benefit of targeting customers from all deciles is further supported by the higher average performance across uplift modeling strategies on the 9th decile. Every pair is capable to increase that number right from the beginning. Only TCIA coupled with SVC fails to achieve this. Second, we now clearly see that ITM coupled with SVC and the two-model uplift method coupled with KNN outperform all other uplift modeling strategies. It can be seen, for example, that both couples perform better than the average performance starting from the 3rd decile. We also note that the performance of the two-model uplift method paired with KNN deteriorates starting from the 7th decile and becomes even lower than the average rate in the 9th decile. This is not valid for the ITM-based SVC. However, we reiterate that ITM-based SVC has shown high variance in the previous analysis (see 6.1). That is, we conclude that there are more signals in favor of the couple of the two-model uplift method and KNN. This couple outperforms all other pairs (including ITM-based SVC) starting from 4th and ending with the 6th deciles. Third, we regard CVT, pessimistic uplift modeling, LGWUM, and LWUM as second-best choice since these strategies perform similar to the average. For example, pessimistic uplift modeling paired with RFC performs slightly better than the average in the first deciles, similar to average in the middle deciles, and underperforms in the last deciles. On the contrary, LGWUM coupled with RFC underperforms the average until the 5th decile and thereafter slightly outperforms the average. Fourth, we observe that combinations of reflective uplift modeling and LogR as well as TCIA and SVC show the weakest performance. Both are clearly inferior to the average. This is especially relevant for the pair of reflective uplift modeling and LogR, since we observe the underperformance in every decile. Thus, we cannot recommend adopting these modeling strategies for similar marketing campaigns. Given that RFC is the best choice in terms of base learners, Fig. 2 suggests that it best performs coupled with the pessimistic uplift modeling strategy since it demonstrates until the 5th decile better or identical performance as average does; this is not given by other combinations.

To get more confidence in the findings obtained so far, we present the *non-cumulative* numbers of incremental purchases in the subsequent experiment. As before, the results are based on the out-of-sample test set. Figure 3 summarizes the respective results for the winner pairs on a decile-level.

Figure 3 provides further findings. Given that truly valuable uplift models are capable to sort customers with high uplift to the first deciles and customers with comparably lower uplift or even negative to latter deciles,[21] we first conclude that CVT, LWUM, and pessimistic uplift modeling perform quite well in the first decile.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

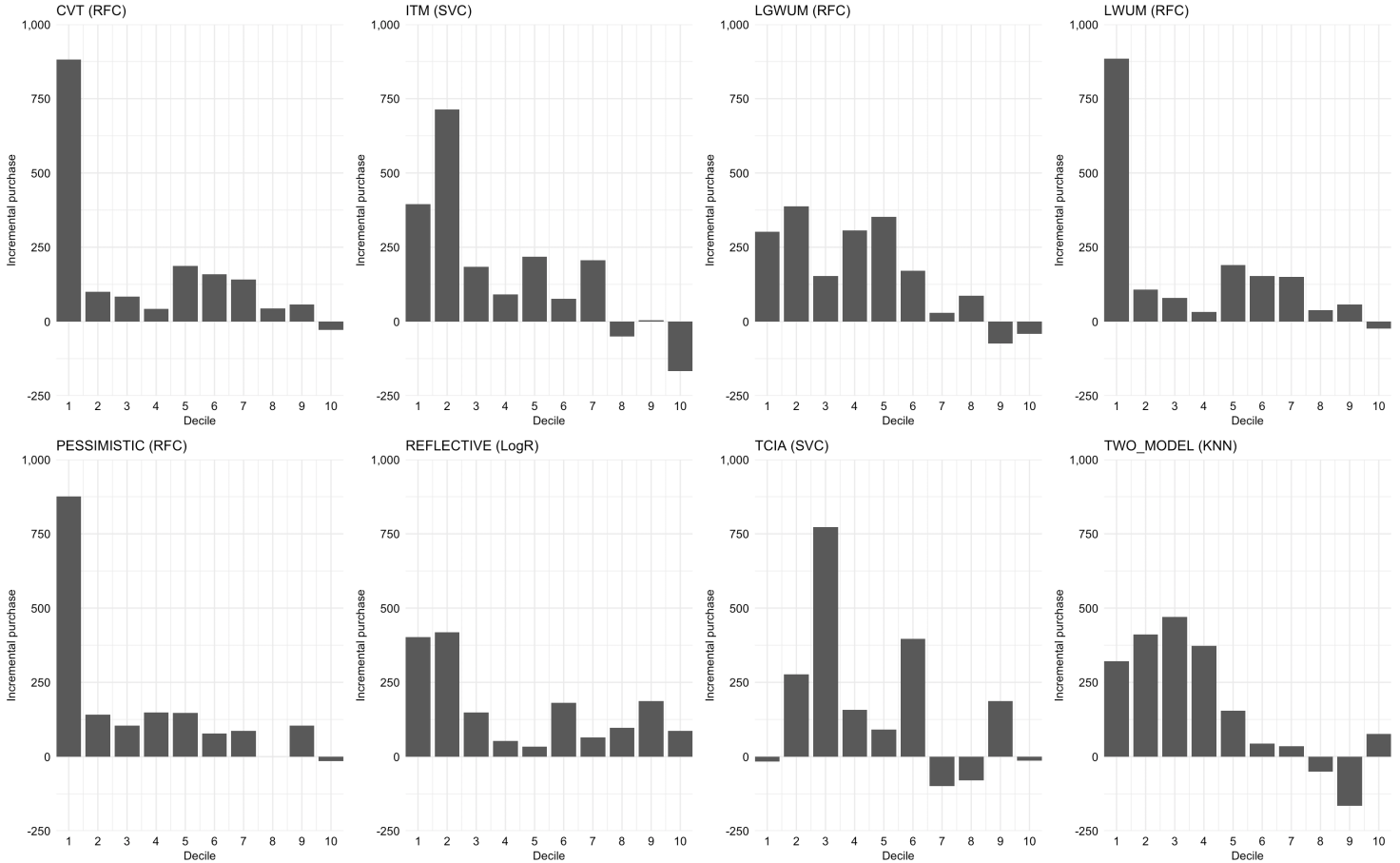

Fig. 3. Non-cumulative numbers of incremental purchases.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Second, comparing the winner pairs as per uplift gain charts — ITM coupled with SVC and the two-model uplift method coupled with KNN — we now are more confident that there are more signals in favor of the latter pair. This is because the two-model uplift method paired with KNN is able to assign customers who are likely to induce positive uplift to the first deciles and negative uplift to the latter deciles gradually. Although ITM-based SVC presents a powerful alternative achieving similar results, we observe that it assigns more customers in the latter deciles than two-model uplift method-based KNN. See, for example, the 5th, 6th, and 7th deciles. Beyond this, we observe that ITM-based SVC assigns less customers in the 4th decile than in the 5th, less in the 6th than in the 7th, indicating unstable results. Third, Fig. 3 provides more confidence in fact that the combinations TCIA and SVC as well as reflective uplift modeling and LogR present the least valuable alternatives. This is because the former allocates customers with negative uplift to the first decile and the latter presents a modeling strategy with no negative uplift in any decile. Given the shortcoming of the pair of TCIA and SVC in the first decile and the fact that it exhibits more variance in the latter deciles, we conclude that this pair presents the worst uplift modeling strategy considered in this study. Therefore, we caution against using this combination.

6.3. *Performance comparison between response and uplift modeling*

Our final experiment is devoted to the examination of the performance of response modeling, a conventional method in marketing applications, *vis-à-vis the best* — two-model uplift method paired with KNN — and *the worst* — TCIA paired with SVC — uplift modeling strategies. To provide a holistic picture on the performance of response modeling, we reiterate all previous experiments, represent the performance of the best and the worst uplift modeling strategies, and extend these experiments by the estimates obtained from response modeling. To secure fair empirical comparisons, we execute response modeling to the same out-of-sample test set for all experiments. We examine the interaction between the modeling strategies and incorporate Q_{pct}. Table 9 mimics the same setup for the interaction examination and adds response modeling to the modeling strategies for conversion uplift (see last row of the table).

Table 9 shows that response modeling outperforms TCIA. That is because it achieves higher Q_{pct} values for two thirds of all base learners (i.e., KNN, LogR, NB, and RFC). Furthermore, we observe that the highest Q_{pct} value of response modeling coupled with RFC is higher than that of TCIA coupled with SVC, 5.679 and 3.893, respectively. This indicates that response modeling might be more beneficial than modern uplift modeling strategies. However, we also see that response modeling fails to outperform the two-model uplift method. Apart from RFC, the two-model uplift method is superior compared to response modeling in every combination. We observe that the interaction of the two-model uplift method with KNN contributes to higher Q_{pct} value than the best combination of response modeling, 7.267 and 5.679,

Table 9. Qini coefficient of selected modeling strategies.

Modeling strategy	Base learner					
	KNN	LogR	NB	SGDC	SVC	RFC
TCIA	1.043	−1.950	−2.821	1.222	**3.893**	0.403
TWO_MODEL	**7.267**	4.305	3.297	0.688	2.806	5.401
RESPONSE	4.752	4.263	0.432	0.546	1.893	**5.679**

respectively. We understand that response modeling interacts best with RFC. This generalizes our finding that RFC is the winning base learner in terms of interaction with uplift modeling strategies. On the contrary, NB and SGDC show worst results when interacting with response modeling; finding that alerts to not execute these base learners for neither uplift nor response modeling.

Figure 4 presents the robustness procedure, aggregation of the results across the 10-fold cross validation, to judge about the variance in the results.

Figure 4 illustrates that response modeling is superior to TCIA since it exhibits smaller variance in the estimates (see, for example, RFC or SGDC) and better interacts with NB and SGDC than TCIA does. We now also see that response modeling interacts with KNN and LogR comparably well to RFC and conclude that the former two base learners are promising when being paired with response modeling. Figure 4 also confirms that response modeling is inferior to the two-model uplift method. We understand that the big share of NB, SGDC, and SVC estimates negative scaled values for Qini, while this is only the case for SGDC when paired with the two-model uplift method (outliers not considered). Moreover, we observe that response modeling interacting with SVC and RFC exhibits higher variance than the two-model uplift method with the same base learners.

We now examine the potential of response modeling to contribute to business value in terms of cumulative and non-cumulative incremental purchases. We echo the same experiments from 6.2 and extend them by the estimates of response modeling. First, we examine the tabular view of the cumulative number of incremental purchases. Recall that figures marked in italic and bold face indicate the same logic as in 6.2.

Table 10 provides evidence on the demerit of targeting the complete customer base compared to targeting a selected share of the population. As in 6.2, targeting the first five to nine deciles is more beneficial than targeting every customer. Apart from this, Table 10 confirms the superiority of response modeling over TCIA in terms of business value. We observe that response modeling holds two global wins, i.e., in the first and in the 9th deciles (i.e., 917 and 1,813 cumulative incremental purchases, respectively), while TCIA has none. However, response modeling is inferior to the two-model uplift method, since the latter holds global wins starting from the 2nd and ending with the 8th deciles. Table 10 also reveals that response modeling interacts successfully with LogR, KNN, and SVC apart from RFC (see number of wins; marked in italic face). Although the pair of response modeling and RFC holds only

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

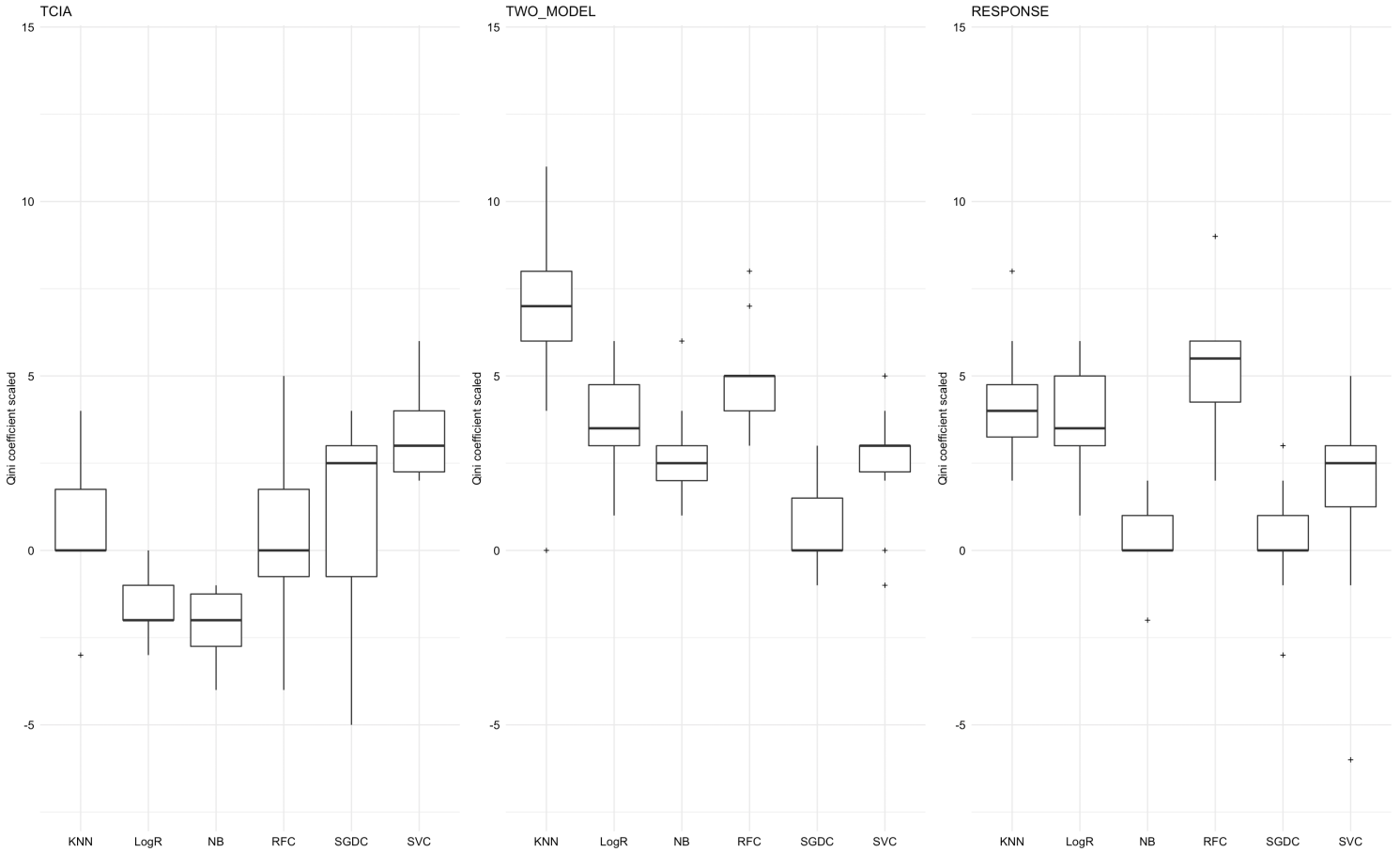

Fig. 4. Scaled Qini of selected modeling strategies.

Table 10. Summary of cumulative number of incremental purchases.

Modeling strategy/ base learner	Cumulative number of incremental purchases per decile									
	1	2	3	4	5	6	7	8	9	10
TCIA										
KNN	133	441	654	795	994	1,220	1,322	1,305	1,371	1,671
LogR	103	60	88	−96	399	962	1,171	*1,573*	1,668	1,671
NB	11	229	84	64	463	710	965	1,221	*1,709*	1,671
SGDC	*309*	*470*	711	843	980	1,065	1,235	1,388	1,582	1,671
SVC	−17	261	*1,033*	*1,190*	*1,281*	*1,677*	*1,578*	1,498	1,685	1,671
RFC	249	423	642	802	961	1,026	1,085	1,157	1,454	1,671
TWO_MODEL										
KNN	321	732	***1,202***	***1,576***	***1,730***	***1,775***	***1,810***	1,760	1,594	1,671
LogR	864	***1,126***	796	1,002	1,257	1,059	1,319	1,503	1,542	1,671
NB	82	583	976	1,183	1,345	1,364	1,410	1,347	1,364	1,671
SGDC	162	421	609	727	929	1,010	1,223	1,378	1,536	1,671
SVC	−49	62	250	900	1,285	1,655	1,785	***1,881***	*1,675*	1,671
RFC	*877*	1,111	1,097	1,117	1,165	1,273	1,437	1,502	1,641	1,671
RESPONSE										
KNN	295	626	933	1,179	*1,417*	*1,509*	*1,612*	1,562	1,641	1,671
LogR	***917***	*1,014*	733	1,202	1,154	1,368	1,287	1,318	1,445	1,671
NB	−206	214	442	623	1,285	1,197	1,349	1,358	1,555	1,671
SGDC	154	388	601	724	946	1,004	1,195	1,363	1,521	1671
SVC	−87	−34	406	784	1,158	1,586	1,507	*1,684*	***1,813***	1,671
RFC	897	853	*1,199*	*1,307*	1,340	1,393	1,457	1,475	1,486	1,671

two wins compared to three wins of the pair of response modeling and KNN, we conclude that the former is the best choice, since this finding is previously supported by the examination of Qini coefficient and robustness procedure. Therefore, we now examine the performance of this best pair compared to the other two best pairs. Recall that TCIA performs best with SVC and the two-model uplift method with KNN. Figure 5 presents related uplift gain charts.

Figure 5 provides new insights. First, we see that response modeling is more successful in the first three deciles compared to the average. Recall that we now average the performance of only these three winner pairs. The performance of response modeling coupled with RFC deteriorates from the 4th decile. The pair TCIA and SVC outperforms response modeling paired with RFC in the latter deciles. See, for example, the 7th, the 8th, and the 9th decile. Figure 5, thus, indicates that TCIA-based SVC might be more beneficial when contacting a larger fraction of customers than response-based RFC. Figure 5 also confirms that the two-model uplift method coupled with KNN is superior over the pair of response modeling and RFC in every decile and achieves higher performance on four deciles compared to targeting the whole customer population.

We now further examine the performance of the winning pairs as per non-cumulative number of incremental purchases. Figure 6 presents the corresponding results in bar charts. Again, we mimic the same experimental setup as in 6.2.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

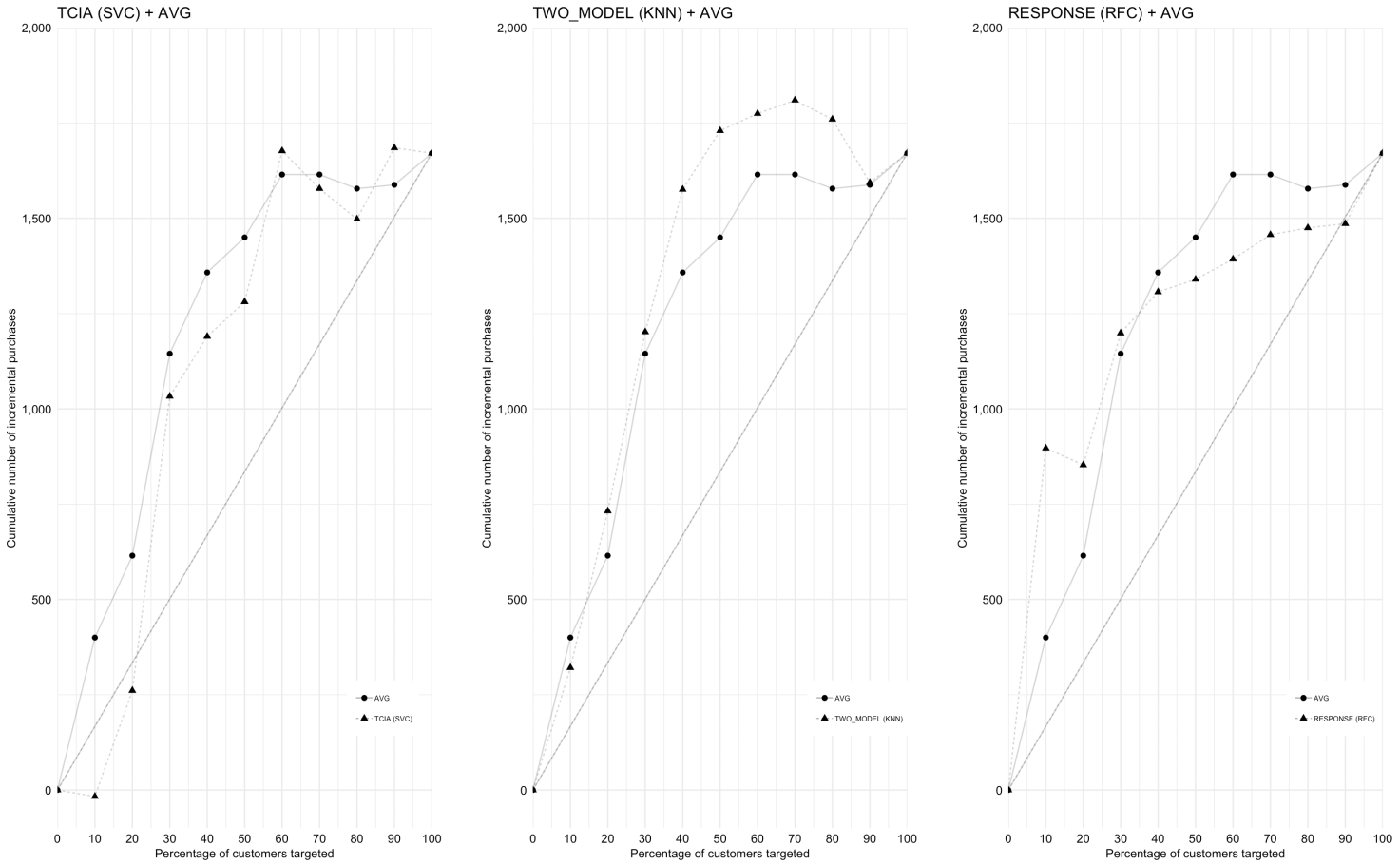

Fig. 5. Uplift gain charts of selected modeling strategies.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

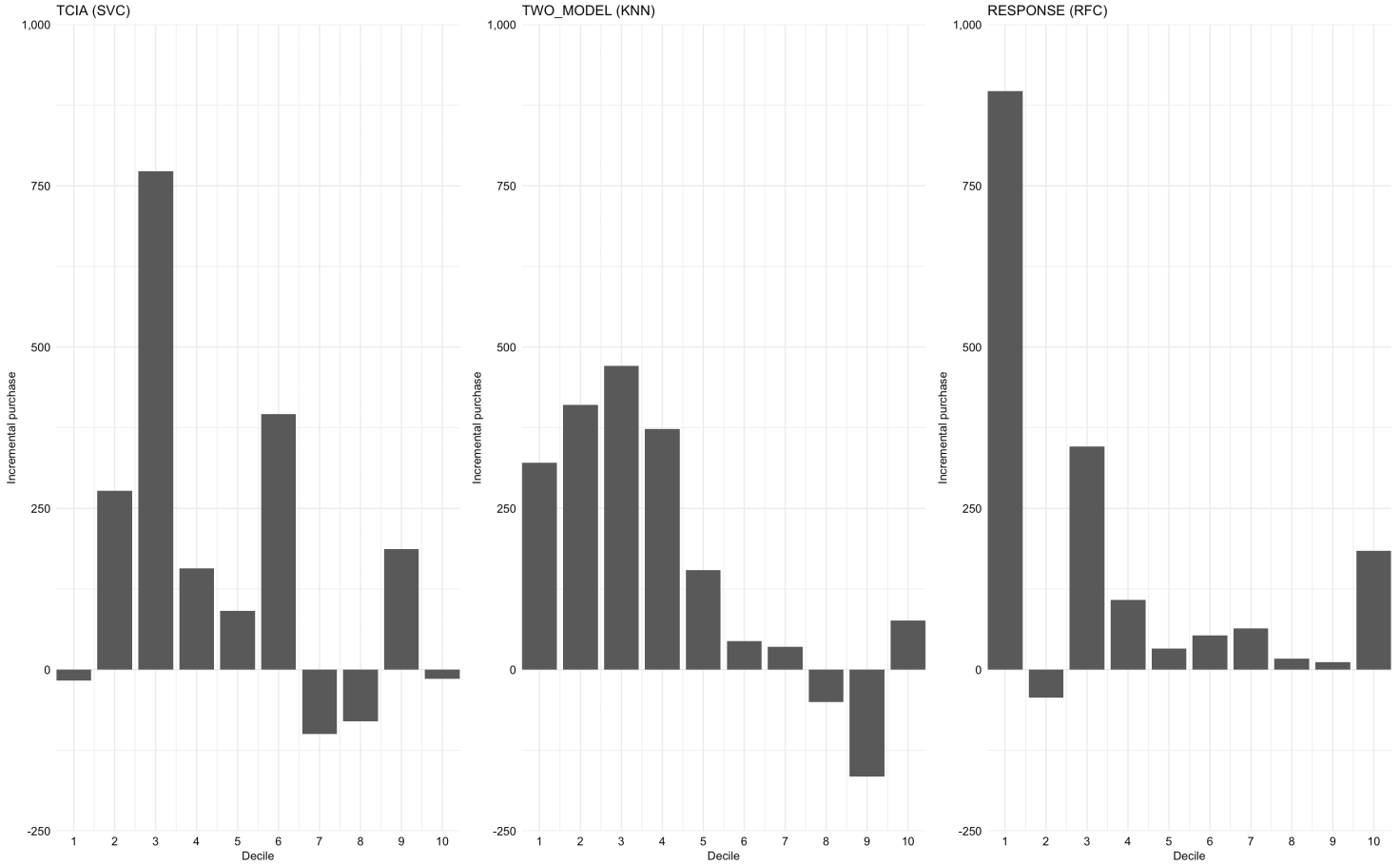

Fig. 6. Non-cumulative number of incremental purchases of selected modeling strategies.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Figure 6 provides the following insights. First, it becomes apparent that response-based RFC performs better than TCIA-based SVC in the first decile. However, we also observe that the former fails to perform in the 2nd decile. Recall that a good strategy aggregates a high number of non-cumulative purchases in the first deciles and small (or even negative) in the latter. Furthermore, response-based RFC fails to assign negative uplift in the latter deciles. We, therefore, conclude that response modeling paired with RFC represents a weak alternative for uplift modeling compared to the two-model uplift method coupled with KNN.

After all, we would like to highlight two fundamental findings. First, response modeling which is usually practiced in marketing applications[9] represents a powerful strategy that leads to success in marketing campaigns as described in this study. We clearly see that it might outperform uplift modeling strategies that have been developed with the purpose to explain the causal relationship between marketing campaigns and an event of interest. And, second, most importantly, that response modeling might be inferior to selected uplift modeling strategies in many experimental dimensions. We, thus, conclude that our study makes it clear that marketers should be aware of the differences among the uplift modeling strategies and apply the best choice in real-world practice.

7. Conclusion

We set out to examine how different modeling strategies for conversion uplift contribute towards increasing the fit of marketing strategies for real-world applications. Uplift modeling can be seen as a technique that patterns causal effect of a marketing incentive on customer behavior. Empirical examination goes alongside multiple dimensions and involves numerous datasets that stem from different geographies and represent distinct product lines. Given that uplift modeling strategies have been proposed in different strands of literature and no attempt has been made to systematically compare predictive performance of them, specific recommendations which strategies achieve highest relative performance have been missing. This study aims to close this research gap through multi-faceted experimentation.

Our study consolidates previous work in conversion uplift and provides a holistic picture of the state-of-the-art in predictive modeling for retail electronic commerce; more specifically, personalized marketing targeting through couponing. From an academic viewpoint, an important question is whether efforts invested to the development of novel uplift algorithms are worthwhile. Our study raises some critical concerns. We find the proposed method to generalize LWUM with weighted probability scores to account for the fraction of treatment and control group customers by Ref. 21 fails to outperform the original LWUM developed by Lai *et al.*[47] in terms of Qini coefficient. A similar picture is obtained in the field of covariates manipulation. We find that TCIA proposed by Tian *et al.*,[34] which to a large extent mimics the procedure of ITM, is inferior to original ITM developed by Ref. 33. On the contrary, we find that ITM, as well as the straightforward two-model uplift method[30] that

captures differences in class probabilities of customers' motivation, represent modeling strategies of first choice for conversion uplift. Our study, therefore, implies that the progress has stalled, and efforts invested to the methodological advancement must be accompanied by a rigorous assessment of new uplift modeling strategies vis-à-vis challenging benchmark. We identify the two-model uplift method and ITM as best performers according to our experiments and advise to compare novel modeling strategies in the field of uplift modeling against them.

An important question to answer in future research concerns the origins of the interaction between uplift modeling strategies and the underlying base learner. We have identified base learners that work specifically well for conversion uplift in digital marketing. However, our study does not seek to explain their success. We believe this is a fruitful avenue for future research; while our study may be regarded as a first move toward gaining insights to this question. For example, we find RFC to interact best with the majority of strategies. This is not given by other base learners. Moreover, RFC performs quite well in the first deciles of targeting and, therefore, can be strongly recommended for campaigns with little budget so that only the 10% most valuable customers are subject to treatment. We find SVC as a valid alternative, although it exhibits high variance in estimates as per robustness procedure presented in this paper. Surprisingly, KNN, usually seen as weak in predictive modeling, has shown appealing results, especially interacting with the two-model uplift method. On the contrary, SGDC and NB have shown poor results in every experiment. We therefore discourage analysts to consider these base learners for uplift modeling.

From a practitioner's viewpoint, it is important to reason whether the observed results can be generalized to real-world applications. On the one hand, we believe that numerous datasets from online shops, several cross-validation repetitions, and performance examination from different perspectives make our results meaningful and relevant for the task of real-time targeting digital coupons in e-commerce. We also believe our main performance criterion, cumulative number of incremental purchases, to approximate the business value of an uplift model, which also raises the relevance of results from a practical point of view. However, we acknowledge that all datasets come from the same provider and exhibit similar features. Uplift models and base learners may behave differently when processing different feature sets. Consequently, we strongly encourage future research to study the behavior of uplift modeling strategies in other marketing and non-marketing applications using different feature sets. For instance, a valuable research avenue might be to extend real-world experimentation by means of publicly available data for uplift estimation (e.g., criteo dataset[c]). Hence, besides the increased model generalizability due to distinct feature sets, researchers might further benefit from reproducibility of results of disclosed data (e.g., Ref. 58).

Apart from this, another suggestion for future research is directed to the introduction of rank-free statistical tests for uplift modeling that provide reliable test

[c]http://ailab.criteo.com/criteo-uplift-prediction-dataset/.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

results due to consideration of effect size even when applying predictive models on large data amounts. Beyond this, study of the influence of uplift models on group decision-making (GDM) processes (e.g., Refs. 77 and 78) as well as on the formation of purchasing decisions through trustful user interactions in online networks (e.g., Refs. 79 and 80) might be fruitful directions for future research. Without claiming external validity, our study may aid corresponding initiatives in pre-selecting promising and less promising modeling strategies.

References

1. K. Bagchi and S. Mukhopadhyay, Predicting global internet growth using augmented diffusion, fuzzy regression and neural network models, *International Journal of Information Technology & Decision Making* **5**(1) (2006) 155–171.
2. Y. Bakos, The emerging role of electronic marketplaces on the Internet, *Communications of the ACM* **41**(8) (1998) 35–42.
3. J. Lee, D.-H. Park and I. Han, The effect of negative online consumer reviews on product attitude: An information processing view, *Electronic Commerce Research and Applications* **7**(3) (2008) 341–352.
4. E. Ascarza, P. S. Fader and B. G. Hardie, *Marketing Models for the Customer-centric Firm*, Handbook of Marketing Decision Models (Springer, New York, 2017), pp. 297–329.
5. T. B. Rhouma and G. Zaccour, Optimal marketing strategies for the acquisition and retention of service subscribers, *Management Science* **64**(6) (2018) 2609–2627.
6. E. Y. Huang and C.-j. Tsui, Assessing customer retention in B2C electronic commerce: An empirical study, *Journal of Marketing Analytics* **4**(4) (2016) 172–185.
7. P. Baecke and D. Van den Poel, Improving purchasing behavior predictions by data augmentation with situational variables, *International Journal of Information Technology & Decision Making* **9**(6) (2010) 853–872.
8. Z. Chen, From data to behavior mining, *International Journal of Information Technology & Decision Making* **5**(4) (2006) 703–711.
9. K. Coussement, P. Harrigan and D. F. Benoit, Improving direct mail targeting through customer response modeling, *Expert Systems With Applications* **42**(22) (2015) 8403–8412.
10. N. S. Sahni, D. Zou and P. K. Chintagunta, Do targeted discount offers serve as advertising? Evidence from 70 field experiments, *Management Science* **63**(8) (2016) 2688–2705.
11. A. W. Ding, S. Li and P. Chatterjee, Learning user real-time intent for optimal dynamic web page transformation, *Information Systems Research* **26**(2) (2015) 339–359.
12. M. Ballings and D. Van den Poel, CRM in social media: Predicting increases in Facebook usage frequency, *European Journal of Operational Research* **244**(1) (2015) 248–260.
13. S. P. Kondareddy, S. Agrawal and S. Shekhar, Incremental response modeling based on segmentation approach using uplift decision trees, *Industrial Conf. Data Mining* (Springer, Berlin, 2016), pp. 54–63.
14. R. Michel, I. Schnakenburg and T. von Martens, Effective customer selection for marketing campaigns based on net scores, *Journal of Research in Interactive Marketing* **11**(1) (2017) 2–15.
15. M. Ieva, F. De Canio and C. Ziliani, Daily deal shoppers: What drives social couponing? *Journal of Retailing and Consumer Services* **40** (2018) 299–303.
16. L. Zhao and J. Zhu, Internet marketing budget allocation from practitioner's perspective, *International Journal of Information Technology & Decision Making* **9**(5) (2010) 779–797.
17. P. Rzepakowski and S. Jaroszewicz, Uplift modeling in direct marketing, *Journal of Telecommunications and Information Technology* **2** (2012b) 43–50.

18. N. J. Radcliffe, Using control groups to target on predicted lift: Building and assessing uplift models, *Direct Marketing Analytics Journal* **3** (2007) 14–21.
19. B. Hansotia and B. Rukstales, Incremental value modeling, *Journal of Interactive Marketing* **16**(3) (2002b) 35–46.
20. S. Jaroszewicz and P. Rzepakowski, Uplift modeling with survival data, *ACM SIGKDD Workshop on Health Informatics (HI KDD'14)*, New York, USA (ACM, 2014).
21. K. Kane, S. Y. V. Lo and J. Zheng, Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods, *Journal of Marketing Analytics* **2**(4) (2014) 218–238.
22. F. Devriendt, D. Moldovan and W. Verbeke, A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics, *Big Data* **6**(1) (2018) 13–41.
23. G. Melli, X. Wu, P. Beinat, F. Bonchi, L. Cao, R. Duan, C. Faloutsos, R. Ghani, B. Kitts, B. Goethals, G. Mclachlan, J. Pei, A. Srivastava and O. Zaïane, Top-10 data mining case studies, *International Journal of Information Technology & Decision Making* **11**(2) (2012) 389–400.
24. D. Zantedeschi, E. M. Feit and E. T. Bradlow, Measuring multichannel advertising response, *Management Science* **63**(8) (2016) 2706–2728.
25. N. Michaelidou and S. Dibb, Using email questionnaires for research: Good practice in tackling non-response, *Journal of Targeting, Measurement and Analysis for Marketing* **14**(4) (2006) 289–296.
26. N. Daskalova, F. R. Bentley and N. Andalibi, It's all about coupons: Exploring coupon use behaviors in email, in *Proc. of the 2017 CHI Conf. on Human Factors in Computing Systems* (ACM, New York, 2017), pp. 1152–1160.
27. S. Lessmann, K. Coussement, K. W. D. Bock and J. Haupt, Targeting customers for profit: An ensemble learning framework to support marketing decision making (2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3130661.
28. V. Chernozhukov, M. Demirer, E. Duflo and I. Fernandez-Val, Generic machine learning inference on heterogenous treatment effects in randomized experiments, *Corr* arXiv:1712.04802v3 (2018).
29. S. Athey and G. W. Imbens, The state of applied econometrics: Causality and policy evaluation, *Journal of Economic Perspectives* **31**(2) (2017) 3–32.
30. N. J. Radcliffe and P. D. Surry, Differential response analysis: Modeling true responses by isolating the effect of a single action, in *Proc. Credit Scoring and Credit Control IV*, Edinburgh, Scotland (Credit Research Centre, University of Edinburgh Management School, 1999).
31. M. Jaskowski and S. Jaroszewicz, Uplift modeling for clinical trial data, *ICML 2012 Workshop on Clinical Data Analysis*, Edinburgh, Scotland (2012).
32. F. Dost, R. Wilken, M. Eisenbeiss and B. Skiera, On the edge of buying: A targeting approach for indecisive buyers based on willingness-to-pay ranges, *Journal of Retailing* **90**(3) (2014) 393–407.
33. V. S. Lo, The true lift model: A novel data mining approach to response modeling in database marketing, *ACM SIGKDD Explorations Newsletter* **4**(2) (2002) 78–86.
34. L. Tian, A. A. Alizadeh, A. J. Gentles and R. Tibshirani, A simple method for estimating interactions between a treatment and a large number of covariates, *Journal of the American Statistical Association* **109**(508) (2014) 1517–1532.
35. L. Guelman, M. Guillén and A. M. Pérez-Marín, Uplift random forests, *Cybernetics and Systems* **46**(3–4) (2015) 230–248.
36. P. Rzepakowski and S. Jaroszewicz, Decision trees for uplift modeling with single and multiple treatments, *Knowledge and Information Systems* **32**(2) (2012a) 303–327.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

37. L. Zaniewicz and S. Jaroszewicz, Support vector machines for uplift modeling, in *Proc. of the 13th IEEE Intern. Cong. on Data Mining Workshops (ICDMW)* (IEEE, Piscataway, 2013), pp. 131–138.
38. S. Hua, What makes underwriting and non-underwriting clients of brokerage firms receive different recommendations? An application of uplift random forest model, *International Journal of Finance & Banking Studies* **5**(3) (2016) 42–56.
39. H. Nassif, F. Kuusisto, E. S. Burnside and J. W. Shavlik, Uplift modeling with ROC: An SRL case study, *Late Breaking Papers of the 23rd Int. Conf. Inductive Logic Programming (ILP'13)*, Rio de Janeiro, Brazil (2013b), pp. 40–45.
40. B. Hansotia and B. Rukstales, Direct marketing for multichannel retailers: Issues, challenges and solutions, *Journal of Database Marketing & Customer Strategy Management* **9**(3) (2002a) 259–266.
41. D. M. Chickering and D. Heckerman, A decision theoretic approach to targeted advertising, in *Proc. of the 16th Conf. on Uncertainty in Artificial Intelligence*, Stanford, CA, USA (Morgan Kaufmann Publishers Inc., 2000), pp. 82–88.
42. N. J. Radcliffe and P. D. Surry (2011). Real-world uplift modelling with significance-based uplift trees. Portrait Technical Report, TR-2011-1.
43. L. Guelman (2014). Optimal personalized treatment learning models with insurance applications, PhD in Economics, Universitat de Barcelona.
44. L. Breiman, Random forests, *Machine Learning* **45**(1) (2001) 5–32.
45. M. Sołtys, S. Jaroszewicz and P. Rzepakowski, Ensemble methods for uplift modeling, *Data Mining and Knowledge Discovery* **29**(6) (2015) 1531–1559.
46. A. Shaar, T. Abdessalem and O. Segard, Pessimistic Uplift Modeling, *CoRR* **abs/1603.09738** (2016).
47. Y.-T. Lai, K. Wang, D. Ling, H. Shi and J. Zhang, Direct marketing when there are voluntary buyers, in *Proc. 6th Int. Conf. Data Mining (ICDM)*, Hong Kong, China (IEEE Computer Society, Washington, DC, USA, 2006), pp. 922–927.
48. C. Manahan, A proportional hazards approach to campaign list selection, in *Proc. of the SAS User Group Intern. Meeting* (2005).
49. K. Larsen. (2010). Net Lift Models. Slides of a talk given at the A2010 — Analytics Conference, September 2–3, Copenhagen, Denmark.
50. S. Jaroszewicz and L. Zaniewicz, *Székely Regularization for Uplift Modeling, Challenges in Computational Statistics and Data Mining*, eds. S. Matwin and J. Mielniczuk (Springer International Publishing, Switzerland, 2016), pp. 135–154.
51. F. Kuusisto, V. S. Costa, H. Nassif, E. Burnside, D. Page and J. Shavlik, *Support Vector Machines for Differential Prediction, Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014*, eds. T. Calders, F. Esposito, E. Hüllermeier and R. Meo, Nancy, France, September 15–19, 2014. Proceedings, Part II (Springer Berlin Heidelberg, Berlin, Heidelberg, 2014), pp. 50–65.
52. T. Cai, L. Tian, P. H. Wong and L. J. Wei, Analysis of randomized comparative clinical trial data for personalized treatment selections, *Biostatistics* **12**(2) (2011) 270–282.
53. L. Guelman, M. Guillén and A. M. Pérez-Marín, *Random Forests for Uplift Modeling: An Insurance Customer Retention Case, Modeling and Simulation in Engineering, Economics and Management: International Conference, MS 2012*, eds. K. J. Engemann, A. M. Gil-Lafuente and J. M. Merigó, New Rochelle, NY, USA, May 30–June 1, 2012. Proceedings (SpringerBerlin Heidelberg, Berlin, Heidelberg, 2012), pp. 123–133.
54. L. Guelman, M. Guillen and A. M. Pérez Marín, Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study, UB Riskcenter Working Paper Series, 2014/06 (2014).

55. B. B. Hansen and J. Bowers, Covariate balance in simple, stratified and clustered comparative studies, *Statistical Science* **23**(2) (2008) 219–236.
56. K. Imai and M. Ratkovic, Estimating treatment effect heterogeneity in randomized program evaluation, *The Annals of Applied Statistics* **7**(1) (2013) 443–470.
57. S. R. Cole and E. A. Stuart, Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 Trial, *American Journal of Epidemiology* **172**(1) (2010) 107–115.
58. K. Hillstrom, MineThatData (2008), https://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html.
59. D. P. Green and A. S. Gerber, *Get Out the Vote: How to Increase Voter Turnout* (Brookings Institution Press, 2015).
60. D. Dheeru and E. Karra Taniskidou, UCI Machine Learning Repository (2017), http://archive.ics.uci.edu/ml/.
61. A. F. J. Connors, T. Speroff, N. V. Dawson, C. Thomas, F. E. H. Jr., D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, W. J. F. Jr., H. Vidaillet, S. Broste, P. Bellamy, J. Lynn and W. A. Knaus, The effectiveness of right heart catheterization in the initial care of critically iii patients, *The Journal of the American Medical Association* **276**(11) (1996) 889–897.
62. M. Pintilie, *Competing Risks: A Practical Perspective* (John Wiley & Sons, Ltd Chichester, 2006).
63. V. S. Lo and D. A. Pachamanova, From predictive uplift modeling to prescriptive uplift analytics: A practical approach to treatment optimization while accounting for estimation risk, *Journal of Marketing Analytics* **3**(2) (2015) 79–95.
64. H. Nassif, F. Kuusisto, E. S. Burnside, D. Page, J. W. Shavlik and V. S. Costa, Score as you lift (SAYL): A statistical relational learning approach to uplift modeling, machine learning and knowledge discovery in databases, *European Conference, ECML/PKDD*, Prague, Czech Republic (Springer, 2013a), pp. 595–611.
65. X. Su, J. Kang, J. Fan, R. A. Levine and X. Yan, Facilitating score and causal inference trees for large observational studies, *Journal of Machine Learning Research* **13** (2012) 2955–2994.
66. F. H.-L. Yong, Quantitative methods for stratified medicine, Doctoral dissertation, Harvard University (2015).
67. H. Nassif, Y. Wu, D. Page and E. S. Burnside, Logical differential prediction bayes net. improving breast cancer diagnosis for older women, *American Medical Informatics Association Symposium (AMIA)*, Chicago (2012), pp. 1330–1339.
68. H. Nassif, D. Page, M. Ayvaci, J. Shavlik and E. S. Burnside, Uncovering age-specific invasive and DCIS breast cancer rules using inductive logic programming, in *Proc. 1st ACM Int. Health Informatics Symp.*, Arlington, Virginia, USA (ACM, 2010), 1883005.
69. E. Braunwald, M. Domanski, S. Fowler, N. Geller, B. Gersh, J. Hsia, M. Pfeffer, M. Rice, Y. Rosenberg and J. Rouleau, Angiotension-coverting-enzyme inhibition in stable coronary artery disease, *New England Journal of Medicine* **351** (2004) 2058–2068.
70. S. Loi, B. Haibe-Kains, C. Desmedt, F. Lallemand, A. M. Tutt, C. Gillet, P. Ellis, A. Harris, J. Bergh, J. A. Foekens, J. G. Klijn, D. Larsimont, M. Buyse, G. Bontempi, M. Delorenzi, M. J. Piccart and C. Sotiriou, Definition of clinically distinct molecular subtypes in estrogen rceptor-positive breast carcinomas through genomic grade, *Journal of Clinical Oncology* **25** (2007) 1239–1246.
71. D. Van den Poel and W. Buckinx, Predicting online-purchasing behaviour, *European Journal of Operational Research* **166**(2) (2005) 557–575.
72. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.

M. Brucher, M. Perrot and É. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12** (2011) 2825–2830.
73. B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens and J. Vanthienen, Benchmarking state-of-the-art classification algorithms for credit scoring, *Journal of the Operational Research Society* **54**(6) (2003) 627–635.
74. S. Lessmann, B. Baesens, H.-V. Seow and L. C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update, *European Journal of Operational Research* **247**(1) (2015) 124–136.
75. J. O. Ogutu, H.-P. Piepho and T. Schulz-Streeck, A comparison of random forests, boosting and support vector machines for genomic selection, *BMC proceedings* (BioMed Central, 2011), p. S11.
76. P. W. Holland, Statistics and causal inference, *Journal of the American Statistical Association* **81**(396) (1986) 945–960.
77. G. Li, G. Kou and Y. Peng, A group decision making model for integrating heterogeneous information, *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **48**(6) (2018) 982–992.
78. J. A. Morente-Molinera, G. Kou, K. Samuylov, R. Ureña and E. Herrera-Viedma, Carrying out consensual group decision making processes under social networks using sentiment analysis over comparative expressions, *Knowledge-Based Systems* **165** (2019) 335–345.
79. R. Ureña, G. Kou, Y. Dong, F. Chiclana and E. Herrera-Viedma, A review on trust propagation and opinion dynamics in social networks and group decision making frameworks, *Information Sciences* **478** (2019) 461–475.
80. Y. Zhao, G. Kou, Y. Peng and Y. Chen, Understanding influence power of opinion leaders in e-commerce networks: An opinion dynamics theory perspective, *Information Sciences* **426** (2018) 131–147.

Int. J. Info. Tech. Dec. Mak. 2019.18:747-791. Downloaded from www.worldscientific.com
by KELDYSH INSTITUTE OF APPLIED MATHEMATICS OF RUSSIAN ACADEMY OF SCIENCES on 11/11/22. Re-use and distribution is strictly not permitted, except for Open Access articles.